Application of Bayesian networks to a longitudinal asthma study

Michael Luke Walker

Master of Philosophy

September 2016

Department of Pathology, MDHS
The University of Melbourne

Submitted in Total Fulfillment of the
Requirements of the Degree of Master of Philosophy

Produced on archival quality paper

Abstract

Asthma is a highly prevalent and often serious condition causing significant illness and sometimes death. It typically consumes between 1-3% of the medical budget in most countries and imposes a disease burden on society comparable to schizophrenia or cirrhosis of the liver. Its causes are as yet unknown but a significant number of risk factors, covering such diverse factors as viral infections during infancy, blood antibody titres, mode of birth and number of siblings have been identified. In recent years there has been increasing recognition of the role played by the microbiome in human health, with a growing understanding that our relationship with the microbes that colonise the different parts of the human body is symbiotic. Disruptions in the microbiome have been implicated in diseases such as obesity, autism and auto-immune diseases, as well as asthma.

At the same time there has been an increasing awareness in asthma research that its multi-facted and multi-factorial nature requires more sophistication than statistical association and regression. In this spirit we employ Bayesian networks, whose properties render them suitable for representing time-direct or even causal relationships, to gain insight into the nature of asthma.

We begin with an example of the simplest Bayesian networks, a linear classifier, with which we predict outcomes in the fifth *year-of-life* according to the statistical distribution of variables from the first two *years-of-life*. (The qualification linear refers to the neglect of correlation and interaction among the predictive variables.)

While classifiers have long been used for prognosis and diagnosis, we use them to

identify useful asthma subtypes, called *endotypes*. Different endotypes often require different treatments and management programs, and driven by different biological factors. These different factors provide different predictors, and a predictor which separates one endotype from the healthy may not do so for a different endotype. We use this to mathematically construct an indicator of when a given predictor is exclusively predictive of a given endotype. Our so-called "exclusivity index" is quantitatively precise, unlike a significance threshold.

The Cohort Asthma Study, whose longitudinal data we analyse, includes the relative abundances of *genera* present in the nasopharyngeal microbiome. In an apparent diversion, we use qq-plots to indicate relationships between the infant microbiome and fifth-year *wheeze*- and *atopy*- status. Interestingly, the relative abundance of *Streptococcus* under certain circumstances was found to be highly predictive of one of the endotypes we identified in the preceding chapter.

Finally, we address the problem of mapping out the complicated interactions among multiple variables. Our model is an adaption of a package originally designed for inferring gene-interaction networks, called *ARTIVA*. This was a non-trivial matter requiring us to augment the discrete data values in order to make them compatibile with the underlying mathematics of *ARTIVA*'s algorithm. With questions from the asthma literature and the posterior probabilities output by *ARTIVA*, we were guided to networks of the interactions between *atopy*, *wheeze* and infection, and could see the difference in the development of immunity-related variables between those who went on to exhibit *wheeze* in the fifth *year-of-life* and those who did not. Our model yielded networks indicating that sensitivity to *viral* infection is an effect and not a cause of *atopy* and *wheeze*.

Statement of Contents

This thesis is submitted for candidature for the degree of Master of Philosophy.

The candidate declares that the contents of this thesis are entirely his own work, except where due reference has been made.

This thesis is less than 50,000 words in length, including all tables and captions.

Acknowledgements

This work was done under the supervision of Associate Professor Michael Tadao Inouye. The candidate also thanks the other members of his thesis committee, Professors Paul Waring, Gary Anderson and Sally Wood, for helpful discussions and other involvement. Helpful discussions with other members of the Inouye Lab are also appreciated. The candidate thanks Dr Gad Abraham, Dr Shu-Mei Teo and Dr Howard Tang in particular. This thesis was written as part of the NHMRC grant APP1049539, and its desired results were part of that project's aims. The candidate also thanks the Department of Pathology (MDHS) at the University of Melbourne for additional support from the Beaney Scholarship for two years from 2013 to 2015.

The candidate's final thanks are to his wife, Francine, and daughter, Yukiko, for their love and support.

Contents

Chapter 1

Introduction

1.1 Relevance and importance

Asthma is currently a serious health issue in Australia and internationally [1–3]. Its prevalence grew rapidly in the latter half of the twentieth century [1,2] and continues to rise rapidly in developing countries although it appears to have peaked elsewhere [1,3].

Approximately 250,000 deaths per year are due to asthma while approximately 60,000 people require hospitalisation each year in Australia alone. The economic burden of asthma in most countries is estimated at 1-3% of the total medical budget [4].

Although the cause of asthma is unknown, many risk factors have been identified. This study has access to a unique, world-class data resource, the Cohort Asthma Study (CAS) [5,6]. CAS is a comprehensive longitudinal five-year cohort study, in which risk-factor relevant data was taken in infancy and then yearly for the first five years of life. The data was taken from children with family histories of asthma or allergic *atopy* [5,6], and includes numbers and types of respiratory infections, immunological factors in the blood, PCR viral data, and environmental factors such as the presence of older children, and exposure to cigarette smoke and common allergens. It also contains sequenced 16S rRNA data from swabs of the nasopharyngeal (NP) microbiome, allowing a detailed picture of which microbes are present. We studied this data at the *genus* level for associations with later *wheeze*-status. In addition to supporting and expanding on a result of Teo *et al.* [7], we also found some new associations between the composition of the NP microbiome and later *wheeze*- and *atopy*- status.

We sought to illuminate the mechanisms of asthma development by identifying time-sequential, and thus potentially causal, links in this data, and thus pave the way for

effective treatment and prevention strategies.

1.2 Conceptual framework

We sought to infer new insight into the nature of asthma from the data in CAS. The data of which we made the most use is related to the number of different types of infections, *IgE* titres, and the presence of *wheeze*, *atopy* and *airborne-atopy*. We also made limited use of interleukin (IL) and interferon (IFN) data. Infections which manifested in the upper respiratory tract but which did not enter the lower respiratory tract were called *upper respiratory infections* (*URI*). To simplify our notation, we also use *URI* to refer to the number of such infections in any given *year-of-life*, the intended meaning being clear from context. Similarly, *LRI* refers to either infections of the lower respiratory tract (*lower respiratory infections*) or to the number of them in any given *year-of-life*. We apply this dual use across all infection types. An *LRI* accompanied by fever is denoted *febrile-LRI*, or by *wheezy-LRI* if it is accompanied by *wheeze*. These last two infection types are considered *severe* and denoted *severe-LRI*. Some *severe-LRI*s were both *febrile* and *wheezy*, but those that were strictly one or the other are termed *purely-febrile-LRI* and *purely-wheezy-LRI*, respectively. *LRI*s which were not *severe* are considered mild and referred to as *mild-LRI*. When discussing infections for which any *virii* were detected, the corresponding *URI*s and *LRI*s are denoted *viral-URI* and *viral-LRI*, respectively. The corresponding terms for when no virus is present are *nonviral-URI* and *nonviral-LRI*, respectively. We also take combinations with fever and *wheeze*, such as *febrile-viral-LRI*, or *purely-wheezy-nonviral-LRI*, and the same notation is applied to specific *virii*, such as Human RhinoVirus (*HRV*) or Respiratory Syncytial Virus (*RSV*).

We have adopted the definition of *atopy* used in CAS [5, 6], where an individual is deemed to be *atopic* to a given allergen if the level of *IgE* antibody in the blood is greater than .35 kilo-units per litre (kU/L). An individual is considered *atopic* if they are *atopic* to one or more allergens. *Airborne-atopy* is *atopy* to an airborne allergen, and an individual is deemed to have been *aeroatopic* if they were *atopic* to at least one

airborne allergen. The various *IgE*-related variables describe the concentration of the corresponding *IgE*-antibody in kU/L, while interleukin data is measured in picograms per mol (pg/mol). A detailed list of variables used, either from CAS directly or constructed from CAS variables, is given in appendix J.

Since asthma is difficult to define or diagnose, we used the presence of *wheeze* as a proxy for asthma, with *wheeze* in the fifth *year-of-life* taken to mean that the child had developed chronic asthma. (Participants in the CAS study were followed up in the tenth *year-of-life*, although the attrition rate was too high for us to find confident results.)

1.3 Outline of thesis

The next chapter contains our reviews of literature relevant to asthma, Bayesian networks and classifiers, and explains both biological and mathematical issues to the level needed to read this work.

In chapter 3 we discerned important predictors and relevant endotypes from the mathematical properties of linear classifiers. The importance of asthmatic endotypes has gained increasingly greater recognition in the literature over the past few years [1–4]. The reasons for this include aetiology, triggers, severity, and medication response, all of which are important for both treatment and research. Crucially, different endotypes also have different risk factors and may be used to help classify them, so an important part of this work was demonstrating that certain variables are predictive only of specific endotypes. We contend that the existence of a predictor exclusive to a given *wheezy* endotype is evidence that that endotype is of biological interest.

This is supplemented with an apparent detour in chapter 4 in which we use dual qq-plots, pairs of qq-plots generated for case and control samples on the same axes, to detect assocations between the *genera* in the infant microbiome and later *atopy*- and *wheeze*- status. Apart from identifying some associations in infectious microbiome samples, we confirm and then explore a recently discovered association [7] between *Streptococcus* in early microbiome samples from individuals who were not suffering a

respiratory infection and later *wheeze*. The extension culminates in finding *Streptococcus*, under these conditions, to be highly predictive of a specific *atopic-wheeze* endotype presented in section 4.3.

In chapter 5, in order to address questions concerning asthma aetiology and other developmental details, we inferred time-ordered, relationships among sequential CAS data. Because of asthma's multi-factorial nature, our approach was to infer networks, specifically Bayesian Networks (BNs). These are directed graphs whose directed edges indicate dependence between nodes, where the nodes represent data fields. To accomodate self-interaction and changes in the network over time we specifically used Dynamic Bayesian Networks (DBNs), in which the network nodes are replicated at every timestep. There are no edges within a timestep, only between subsequent time steps. This allows a DBN to represent self-interaction by an edge between corresponding nodes in adjacent timesteps. Indirect self-interaction appears as a directed path starting from a node, passing through one or more non-corresponding nodes over one or more time steps, and then returning to the corresponding node at a later timestep. We give a detailed introduction to DBNs in section 2.2. Our main technical obstacle in this chapter is that the package (*ARTIVA*) we used to infer the DBNs is inherently unsuitable for binary and count data, such as the presence or absence of *wheeze* and the number of various types of infection. We surmounted this problem with a code module which augments integer-valued variables with randomly generated continuous variables. Our chief finding in chapter 5 is that infections, even *viral* ones, are subsequent to, and causal of, *wheeze* and *atopy*.

We generated the DBNs using the R-package "ARTIVA". One of the significant advantages of ARTIVA is that it allows the edges to change over time, with a flexible number of change points. Its authors could then apply it to the transition between different phases in the cell growth cycle, and we used it to find critical time points in asthma susceptibility. For example, in subsection 5.5.4 we find that the relationship between *viral-LRI* and the number of allergens to which each participant is *atopic* apparently reverses during the first two *years-of-life*.

Despite its virtues, *ARTIVA* required modification in order to be suitable for our needs. Specifically, it was only suitable for continuous data as its algorithm assumes the model parameters to have a Gaussian distribution in order to simplify the underlying mathematics. (This is described in mathematical detail in the supplement of the original *ARTIVA* paper [8].) Such an assumption is logically inconsistent with the use of discrete variables, such as the binary and count variables common in CAS. This problem was solved by replacing binary and count variables with continuous latent variables, in accordance with the long-known method of Albert and Chib [9].

Bayes' formula indicates that posterior probability is a measure of how well a particular model fits the data. Indeed, with neutral priors the two are proportional. Hence a relationship between data nodes with intermediate posterior probability suggests that it is relevant, but with room for improvement. A mediocre fit has several possible causes, such as nonlinearity and statistical noise, or it could indicate a sub-optimal choice of data field. For example, a highly probable relationship involving *wheezy-LRI* should still fit *severe-LRI* (*wheezy-* and *febrile- LRI*) quite well since *wheezy-LRI* is a significant subset of *severe-LRI*. But if the relationship did not hold at all for *febrile-LRI* then the fit for *severe-LRI* would be inferior to that for *wheezy-LRI*. So, given a DBN involving an intermediate posterior probability involving *severe-LRI*, we tried replacing *severe-LRI* with *wheezy-LRI* or *febrile-LRI* to see if it yields a better fit. In this example, *wheezy-LRI* would have given a better fit while *febrile-LRI* would have given a poorer one.

Chapter 2

Literature Review

2.1 Asthma pathogenesis

Asthma development is believed to entail a complex interplay between genetic and hereditary factors [10]. Furthermore, its manifestation varies greatly over a range of observables, such as severity [11], asthmatic stimuli [10, 12, 13] and age-of-onset [14, 15]. While many correlations have been identified in the literature, it is not yet understood which are of direct relevance or form part of any causal relationship. Although it manifests in the lungs, the pathology derives from an inappropriate immune response [12, 16, 17], which may or may not be *atopic* in nature.

2.1.1 What is asthma?

Asthma is a complicated disease in which the lungs become hypersensitive to various, typically harmless, stimuli. This leads to an excessive immune response which includes airway constriction [12, 16, 17]. The strength of this constriction can range from mild difficulty breathing to a potentially fatal airway restriction requiring strong medication and hospitalization.

The main clinical symptoms of asthma are shortness of breath, coughing and wheezing, usually in response to specific stimuli such as allergens, infection or exercise [12, 16, 18]. Of course, these stimuli are not specific to asthma, and other factors such as family history, recurrent bronchitis and response to certain medications might also be considered.

Asthma displays a variety of phenotypes. One classification scheme [10, 12] describes *atopic*, *non-atopic* and *transient*. *Atopic* asthma is the "classic" asthma phenotype [10, 16], making $\approx 91\%$ of asthma cases according to some sources [11]. It is accompanied by

atopic sensitization before the age of three, with attacks triggered by the corresponding allergens. This is the most common form of asthma, with the biggest risk factors for asthma development being aeroallergen sensitization [16], and early-onset *atopy* [10, 19, 20].

However not all asthma is accompanied by *atopy*, as demonstrated by a study [13] finding that most, but not all, teenage sufferers of asthma are also atopic, while only a minority of *atopy* sufferers have asthma. Asthma without *atopy* accounted for 19% of subjects according to one study [19]. While clearly a minority, it is still a significant minority and one with more severe outcomes [11], including greater loss of lung function [10, 12, 21] and mortality in later life [21]. This last point is well illustrated by *non-atopic* ("intrinsic" in their terminology) asthmatics being over-represented (117 out of 170 asthmatics) in [21], whose subjects were all recruited from the State University Hospital in Copenhagen. In a ten year follow-up study on intrinsic childhood asthma [22] furthermore, nearly one third (24 from 70) of the asthmatics were considered intrinsic, and 20 of those had current symptoms at the ten year follow-up. (The corresponding proportions were similar, 40 from 46, for the *atopic* asthmatics.) Importantly, the outcomes for each of the two asthma types had different predictors, indicating that they have different underlying mechanisms.

The remaining asthma-related phenotype is *transient wheeze*. It is characterised by a *wheeze* in the first three to five years of life, which is no longer present beyond five years [10]. This relatively common [20] phenotype was found to typically demonstrate a reduced lung function prior to infection or other insult, and the lungs continued to develop along a normal trajectory without attaining full normal function.

Another proposed scheme [15] classifies asthma phenotypes according to age-of-onset and the concentration of eosinophils (a type of inflammatory cell). It is based on a study in which early onset (before age 12) wheezers show greater prevalence of allergic triggers such as dust or pollen, but no difference in non-allergic triggers such as tobacco smoke and cold air, while late onset (12 or older) wheezers demonstrate lower lung function [3, 11, 15, 21] despite the absence of significant remodelling [15]. (Remodelling was evaluated

according to the measured subepithelial basement membrane thickness.) The relevance of age-of-onset is supported by a study [14] of eight genes, statisticially associated with asthma before GWAS studies, but not found significant in GWAS studies. The study identified three SNPs which became statistically significant when age-of-onset was taken into account. Similarly, a Taiwanese study [23] found that different genes associated with elevated *IgE* levels, a risk factor for asthma, were associated with different stages of development.

Eosinophil data indicates [15] that elevated eosinophil levels, (defined as more than two standard deviations above the mean), are associated with stronger symptoms, in both early and late onset asthmatics. There is also a pronounced Th2 (allergy-prone) pattern of immune cell expression and inflammation pattern associated with eosinophila among the early onset group but not among the late onset group.

Both early- and late- onset wheezers had both high and low eosinophil subgroups, but high eosinophil levels were more common among the early onset phenotype. There is also evidence [3,15] that severe airway restriction can occur in the absence of apparent inflammation, suggesting a distinct "steroid-resistant" severe asthma phenotype.

A cluster analysis [12] found five clusters with different mixtures of age-of-onset, severity, atopic status and baseline lung function. This further emphasises the complex heterogeneity of asthma.

2.1.2 Epidemiology

Asthma, in all its forms, is a significant global health issue. It is currently estimated that as many as 300 million people have asthma [1–4]. Asthma prevalence rose dramatically over the latter half of the twentieth century [1, 2]. It continues to rise in developing countries although developed countries, with higher asthma incidence, have seen its prevalence level off in more recent years [1,3]. Prevalence in Oceania, among the highest worldwide, has even dropped slightly (0.39%) [24]. In most countries the economic burden of asthma is about 1-3% of the total medical budget [4].

Asthma prevalence in Australia, for reasons unknown, is unusually high by world

standards [1,2]. Australia's asthma prevalence in various age-groups is about 14-15% [1] with 27% of Australian children having experienced asthma symptoms by seven years of age [24].

Asthma is potentially fatal, responsible for approximately one in every 250 deaths [1], about 250000 deaths each year worldwide [4]. In Australia, 60000 people require hospitalisation annually [1]. Worldwide around 15 million Daily Adjusted Life Years [4], about 1% of the total, are lost each year. This is similar to diabetes, cirrhosis of the liver, or schizophrenia [4]. Many are required to manage their asthma with medication regimes, which are often unavailable in under-developed countries [4].

2.1.3 Clinical presentation

The clinical symptoms of asthma arise from an exaggeration of three, normally protective, physiological processes within the lung. The first is that the airways contract due to tightening of the smooth muscles that line them, narrowing the airways and restricting the flow of air. The second is the secretion of mucous, which further narrows the airways with the additional risk of plugging them. The third process is inflammation, a defense against infection, where the lining of the lung becomes swollen and inflamed, narrowing the airways further.

In addition to the remarkable increase in degree, these processes also occur much more readily in the asthmatic lung, typically in reponse to innocuous stimuli. The changes in lung physiology [16, 25–27] contributing to unnecessary sensitivity, excessive airway constriction, and excessive mucous production are excessive innervation, smooth muscle thickening and an increased number of the goblet cells responsible for mucous production. (In chronic asthma the goblet cells spread to the peripheral airways where they normally do not exist [26].) Other changes include the growth of additional small blood vessels (angiogenesis), and thickening of the membrane [16, 25–27] under the epithelial cells lining the airways, causing the lung to lose elasticity and full function. Some authors have suggested that this is a compensatory mechanism to reduce the excessive airway constriction [25–27].

2.1.4 Role of immune system

The remaining asthma pathology is in the immune system, specifically in the relative proportions of the different immune cell types [15,16,28]. Human immunity is comprised [29] of two functionally distinct parts, *innate* immunity and *adaptive* immunity. Innate immunity acts as a first line of defense [29,30] against a broad range of foreign organisms, both through macrophages which attack in response to a limited range of conserved surface molecules, and through inflammation, mediated by leukocytes, of which there are several types. Meanwhile, dendritic cells present antigens specific to the infectious agent to the cells of the adaptive immune system.

The adaptive immune system recognises invasive agents that it has encountered previously and mounts a strong and rapid response when the same pathogen attacks again [29]. The major types include B-cells and T-cells [16, 28, 29]. B cells attack antigens in the blood stream with specific antibodies while T-cells destroy infected cells and their contents. Both T- and B- cells require priming to a particular antigen before they mature and become active. This occurs when the immature T- or B- cell is presented with an antigen by another cell, typically a dendrite [31], a mast cell [32], or a mature T- or B- cell. Some T-cells, called helper T-cells or Th cells, play a controlling role, coordinating the immune response. Th1 [30, 33] and Th17 [34] push the immune system towards an inflammatory response, where Th17 has been implicated in auto-immune diseases [28, 34]. Th2 favours an adaptive, antibody centred response and is widely implicated in allergic disease [28, 33], in which the immune system becomes primed to act against a wide range of innocuous substances such as dust, pollen and other sources of allergy such as cat dander.

The allergic response is a central feature of asthma. Genetic predisposition for allergy, specifically for high *IgE* antibody production in response to allergens, called *atopy*, is a major risk factor for asthma [20], especially when the onset is early [6,35–37]. As an illustration of the strength of the genetic overlap between asthma and other immunity-related disorders, genetic experiments in mice show that immune deficiencies

can reduce susceptibility to asthma risk factors such as fungi [38] and cigarette smoke [39]. Other genetic associations show correlation, others anti-correlation, with auto-immune disease [40].

It is well-known that common allergens often trigger asthma exacerbations (attacks), and atopic indicators such as blood concentrations of *IgE* antibodies are risk factors for asthma. This is well illustrated by one study [19] showing that *IgE* levels at age two is a reliable predictor for persistent *atopy* and wheeze, while another study [41] has found that house dust mite exposure alters the expression of the immune system's T-cells away from the inflammatory Th1 response towards an antibody-oriented Th2 response. There are even genetic factors in common to both asthma and *atopy* [42].

Allergy symptoms can occur in any part of the body, and it is argued [16, 43] that an allergic disorder in one organ, such as dermatitis (eczema) in the skin, can lead to allergic disorders elsewhere, such as the lungs. This "spillover" is thought to result from dendritic cells in all tissues originating in the same compartment of bone marrow. Hence sensitivity to air-borne allergens occurs first [16, 44], leading to asthmatic symptoms later. Atopic march, early dermatitis followed by food allergy and allergic rhinitis (hay fever) [44] is a well-recognised progression through atopic conditions to asthma probably due to this mechanism. A similar model in which atopic dermatitus can lead to asthma has been reported in mice [45, 46], and recent evidence suggests that this might also occur via bone marrow [47].

2.1.5 Role of early-childhood viral *infections*

However atopic asthma, even in early onset asthma [15], is only one phenotype of asthma as discussed earlier. One particularly important *non-atopic* asthma phenotype is virally-induced asthma, which Strippoli *et al.*. [48] concluded was a separate phenotype from *atopic* asthma. A relationship between early respiratory *viral* infection and asthma development was considered over 30 years ago [49] and subsequently indicated by long-term birth cohort studies [20] in Australia [19, 50] and overseas [51, 52].

Probably because it is the most common serious respiratory infection in newborns

[53–55], *viral* studies of asthma pathogenesis are primarily focused on Respiratory Syncitial Virus (RSV), for which associations have been made [56–58], and contradicted [59, 60] with recurrent wheeze and later asthma and *atopy* (all studies found an association with transient wheezing). An analysis by Wennergren and Kristjánsson [60] found that severe RSV followed by wheeze indicated an underlying mutual cause.

There is now strong evidence that Human RhinoVirus (HRV) is also an important factor, with several studies finding HRV to have a strong association with asthma exacerbation [6,61–63]. (RSV and corona virus have also been found to contribute to asthma exacerbation [61].) Furthermore, a lower respiratory infection secondary to HRV in children is a risk factor for conditions such as asthma, other cardiorespiratory problems [64] and cystic fibrosis [65].

One of the biggest risk factors for ongoing *wheeze* and *atopy* is *viral* infection in the presence of early *atopy* [19, 20, 50, 66, 67]. The conventional view is that susceptibility to *viral* infection shares a common cause with early *atopic* sensitization, but an emerging view [6,16] is that *viral* infections can exacerbate an *atopic* condition and even help its establishment in the lungs. The details are not understood, but cross-linking mechanisms, similar to those responsible for the atopic march, between *atopy* and *viral* infection have been proposed [6, 16, 55, 68–72]. One of our more important findings is that this emerging view is not supported by our model.

The central issue is how an anti-*viral* Th1 response can lead to a Th2 allergic response. It has been observed that *viral* lung infections increase the population of various dendrites [73, 74] and other cells concerned with directing the immune response against allergens in the lungs, and that some of these numbers remain high post-infection. Hence an existing *atopy* in another tissue is assisted into the lungs.

There is also evidence that the dendrites themselves are altered, as some *viral* infections up-regulate a key *IgE* receptor (FcϵR1) [16,75], so that Th2 anti-*viral* inflammation is effectively recruited to the already existing allergic response, and also furthers allergic sensitivity via the bone marrow crossover discussed earlier. Such *viral* exacerbations also damage the epithelial lining of the lung, further degrading its barrier function and

leaving it more susceptible to later exacerbations. This closes a vicious circle of inflammation, remodelling and loss of lung function, and enhanced hypersensitivity [16], irreversible once pathological damage has passed a critical level.

2.1.6 Nasopharyngeal microbiome

Bacteria are also relevant to asthma development, though in a different manner to *virii*. It is now known that the human body harbours various bacterial microbe colonies [76–78], and their role in immunity-related disease is an emerging research area [79,80]. The role of microbiota in health generally is still being explored with one recent study [81] showing a link between the gut microbe colony (microbiome) and obesity in mice. Another has found that a "healthy" microbiome can provide resistance against intestinal pathogens [80]. Fujimara *et al.* [82] found that exposing mice to dust from dogs alters their gut microbiome and leaves them less vulnerable to airway insult and allergy. It has also been found [83] that alterations of healthy gut bacteria, especially *Bifidobacterium* and *Lactobacillus*, is a high-risk factor for allergy and asthma. A result in mice, possibly applicable to humans, is that an unhealthy gut microbiome can even contribute to autism [84]. Asthmatics have higher levels [16] of bacteria diversity and density in the nasopharynx (upper part of the airway above the soft palate) although the diversity of this particular microbiome varies widely even among healthy people [76].

While all this suggests a complex relationship between the microbiome and health, a reasonable hypothesis is that the nasopharyngeal microbiome could be a source of opportunistic infection [16, 85], especially given that early microbial colonization is a risk factor for childhood asthma [86].

It could also be seen as an indicator of a defective immune system [16, 85], with reduced capacity to clear bacterial and, perhaps, *viral* incursion. This is supported by findings that children who develop early asthma and *atopy* have lower serum concentrations of bacterial-specific $IgG1$ [87,88], as do children susceptible to virally-triggered asthma exacerbations [89]. Further evidence is that bacteria-specific IgE concentrations increase after virally induced asthma exacerbations [88], consistent with additional ex-

posure to a bacterial antigen. Finally, *IgE* antibodies to *Haemophilus* and *Streptococcus* are protective against asthma [88, 90], probably because they are of a relatively soluble form unsuitable for cross-linking with *atopy via* the FcεR1 receptor discussed above.

Hence the nasopharyngeal microbiome potentially influences asthma development in two ways. The first is by influencing the occurence of the allergic response via interaction with the immune system [83, 90, 91], and the second is as a source of opportunistic bacterial infection [85].

2.1.7 Summary

Asthma is a complex disease involving complicated interplay among the lungs, various facets of the immune system [45–47, 69], the nasopharyngeal [85, 86, 92] and other [7, 93–97] microbiomes, and multiple environmental factors including *viral* infection [19, 48, 51], allergen [41, 44] and cigarette smoke [39] exposure, and even exclusive breastfeeding [50]. To properly demonstrate its origin first of all requires detailed data from deeply phenotyped cohort models, such as the CAS study in Perth [5, 6], the URECA study in Wisconsin, USA [98] and the MAS study in Germany [51], combined with leading edge statistical methods.

2.2 Literature Review - Bayesian Analysis

2.2.1 Bayesian networks

Data obtained from studies such as the longitudinal cohort study CAS [5, 6] are conventionally analysed using standard statistical methods [19] to find associations, covariances and (usually linear) regressive relationships. While these have undoubtedly proven valuable [5, 6, 13, 50], there are risks applying a correlation without elucidating the causal mechanism. For example, a recent publication argued [99] that saturated fat and cholesterol are *not* causal of heart disease, despite their long-accepted association, but may in fact be protective against the stresses that do cause it. If this is so then the medically advised reduction of fat and cholesterol in western diets may have been counter-productive. Conventional statistics can rarely derive more than conditional dependencies without additional input such as intuitive insight or time-sequence data,

and these fall short of causal relationships in that different causal relationships can have matching conditional dependencies [100, 101].

Analysis of a causal system is greatly assisted by representing it on a *Directed Acyclic Graph* (DAG). A *graph* is a set of nodes, connected by a set of edges. If the edges have a direction assigned to them then they are called *directed edges*, running from the *parent*, or *source* node to the *child*, or target node, and the graph is called a *directed graph*. Otherwise they are *undirected*. The correlations found by conventional statistics can be represented by an undirected graph if the nodes are made to correspond to the factors being considered. To represent causal relationships it is necessary to use *acyclic* directed graphs[1]. A graph is *acyclic* if it is not possible to travel along the edges in the graph from a node and return to it without traversing any edge more than once. Otherwise it is called a *cyclic* graph and is said to contain *loops*. A directed graph is cyclic only if a loop can be traversed in the direction assigned to its edges.

While directed edges have an obvious relevance to causal influence, the unidirectional nature of causality requires a directed graph representing causal relationships to be acyclic. A Bayesian Network (BN) [102, 103] is a DAG representing causal relationships among statistical data in which nodes also contain the relevant probabilities given the values of immediately upstream, or *parent*, nodes.

To find a useful BN, especially among a large number of diverse data fields, requires a powerful inferencing technique capable of processing large amounts of raw data. Several machine learning techniques for this are known [104], such as Inferred Causation [100], minimum message length [105], causal minimal message length [106], and various other score-based algorithms [107, 108]. Kidd *et al.* [109] have raised similar issues, arguing for the application of networks, Bayesian methods and information theory in their study of the immune system.

The simplest Bayesian network is the linear classifier, in which the category being determined, called the *response* or *response variable*, is the parent node to every other

[1] We use the term *causal* in the rigorous sense requiring *cause* to precede *effect*, as distinct from simply indicating influence.

variable in the network, called *predictors* or *predictor variables*. We describe it further in section 2.3. For now we continue with the tools needed to infer networks.

2.2.2 Inferring parameters through Bayesian classification

Bayesian analysis is based on Bayes' equation [110],

$$P(\Theta = \theta|D) = \frac{P(D|\Theta = \theta)P(\Theta = \theta)}{\int d\theta' P(D|\Theta = \theta')P(\Theta = \theta')}, \tag{2.1}$$

which gives the *posterior probability* that parameters Θ have value θ given the data D and a *prior expectation* $P(\Theta = \theta)$ (We follow the convention of referring to parameters with upper-case (Θ) letters and their possible values with lower-case (θ)). This *prior*, $P(\Theta = \theta)$, can be interpreted as prior knowledge or belief, and has historically [111] been a source of controversy regarding the role of *belief* and *uninformed priors*. The remaining term in the numerator, $P(D|\theta)$, is the *marginal probability* that the data would be D if $\Theta = \theta$, while the denominator integrates over all possibilities to ensure that the probabilities add up to unity.

We have chosen to use Bayesian inference for its multiple advantages. It is exceptionally tolerance of missing data [112]. Also, it can formally incorporate known information by means of a suitably chosen prior distribution or *prior*. Finally, the Bayesian approach is theoretically immune to irrelevant data [113] and can therefore be used to identify which data is relevant. For example, Christensen *et al.* [114] applied Bayesian methods to selecting among generalizations of the Dirichlet distribution (Polya trees) and notes the flexibility provided by Bayesian methods with the desirable feature that additional data does not complicate a model unless it is truly relevant. We should however note that implemented algorithms can sometimes fall short of this theoretical ideal [113, 115, 116].

The chief Bayesian weakness is vulnerability to correlations among the data fields. Sometimes these correlations can be very detrimental [117]. Attempted remedies include propensity scores [118–120], and combinations [121] of Tree Augmented Näive Bayes (TANs) [122, 123], which add links among data fields to weaken the assumption of their

independence, and greedy [117] search algorithms.

Bayesian vulnerability to correlated data has remarkably little impact on classification problems [124–126], and these were the focus of early Bayesian applications [123, 127, 128]. Indeed, Bayesian classification was applied such diverse topics as quality of care assessment for hospitals [129] and information retrieval [130]. Later came prediction of outcomes of medical and other procedures [129, 131, 132] and the association of genetic data with phenotypes [133]. These methods give the algorithm, usually constructed from data, an item's probability of a certain trait depending on its category. These probabilities are then fed into Bayes' equation, eq. (2.1).

We shall use the simplest and most commonly assumed relationship on a Bayesian network, linear regression [134], where one seeks to infer the linear coefficient. Examples include Yan *et al.* [133] who sought associations between SNPs and quantitative traits. Also, Rigaux *et al.* [135] used Bayesian regression to assess the risk of bacterial food contamination. Taking the standard Quantititative Microbial Risk Assessment (QMRA) model as its prior, the analysis updated about 25% of its parameters, providing evidence against the QMRA model. (Bayesian analysis has also been applied to nonlinear functions [136, 137].)

2.2.3 Implementation

The computationally difficult part of Bayesian analysis is the integral over parameter space (the denominator in eq. (2.1)). Techniques for evaluating this integral form the bulk of research effort in this field. The generally preferred method of evaluating the integral is the Markov Chain Monte-Carlo (MCMC) technique [103, 138–141]. Monte-Carlo is the technique of estimating an integral based on randomly selected points. A Markov chain is usually used to generate the sequence of points, and its defining feature is that the choice of subsequent point is independent of previous points, although it may depend upon the current one. (If its dependence on the current point is sufficiently small it is simpler to implement Tierney's independence sampler, which assumes independence from the current point [142].) If allowed to run for a sufficient period of time

(called the *burn in period*), the chain will display a time-independent distribution over parameter space. The simplest form of MCMC is the Gibb's sampler [103, 140] , in which subsequent points are chosen according to the distribution $P(\theta')$ under the integral. Although it can be applied to multi-variate problems [143], the Gibb's sampler was found to only be well-suited to univariate integrals, and poorly suited for meaningful model comparisons.

These limitations were overcome by generalisation of the Gibb's sampler to the Metropolis-Hastings algorithm [103, 138, 139, 141], in which points are proposed and then accepted or declined with a probability dependent on the marginal likelihood.

The Markov chain is not obligatory. Monte-Carlo methods can be combined [129] with various stochastic methods including simulated annealing and genetic algorithms. Carbonetto and Stephens [144] used a combination of variational methods [145–147] and importance sampling [141] to a Genome Wide Association Study (GWAS) because the calculation time was much shorter than an MCMC calculation would have been. However the Markov chain, being widely applicable and computationally efficient, is by far the most popular choice [141] .

In the examples so far, the network topology is given. Unknown network topologies are often derived by non-Bayesian methods. Such algorithms typically compare alternative networks using so-called *Bayes' factors*. The Bayes' factor is the ratio between the posterior probability of a result and its prior probability. For example, Yu *et al.* [148] generated gene networks, comparing the performance of three optimization algorithms, greedy, simulated annealing, and genetic (the best performer was, due to its speed, greedy with random restarts).

This reflected a general limitation that MCMC could only select values within a given model but not yet choose between different models. This was due to different models not having corresponding parameters or even the same number of degrees of freedom in parameter space, so that the probability of moves in the Markov chain were not necessarily independent of direction. Though not strictly essential, conserved degrees of freedom ensure that the Metropolis-Hastings algorithm generates an equilibrium

distribution of the Markov chain [138, 139, 149].

An early successful application of MCMC to network inference ironically serves to illustrate this limitation. Tali and Hengartner [150] were able to infer a sequence of graphs with MCMC only after specifying the number of graphs in the sequence.

The problem of sampling parameter subspaces of different dimensionality was solved [149] by the *reverse jump MCMC* (RJ-MCMC) method, in which fictitious degrees of freedom are randomly generated in transitions involving a change in dimension. A good description of RJ-MCMC is given by Pagel and Meade [151], which applies it to the coevolution of multiple male mating and oestrus advertising in primates, where it is used to infer the branches of a phylogenetic tree. (Also see the discussion in Dellaportas *et al.* [143].)

While RJ-MCMC is ideal for inferring graphs, the Bayesian networks described so far are not suitable for genetic and other biological networks, for two reasons. The first is that such networks can change their topology over time, such as the network controlling cell growth, in which gene activity and network topology vary with the phase of the growth cycle [8, 152, 153]. The second is that they typically contain self-interaction, either directly as in the case of a gene regulating its own expression, or indirectly as when a species limits its own population growth [136] by depleting its food supply. Such self-interaction, also common in genetic networks [154], corresponds to a directed loop, forbidden in Bayesian networks.

2.2.4 Dynamic Bayesian networks

The solution to both these problems is to replicate the nodes of a Bayesian network at each of multiple time-points [8, 148, 155, 156]. Edges are only allowed from the nodes in one time-point to those in the next one. Network changes correspond to changes in the pattern of edges at corresponding time points, and self-interaction is indicated by directed paths between corresponding nodes. This construction is called a Dynamic Bayesian Network (DBN), and is a Bayesian network in its own right.

Robinson and Hartemink [152, 153] inferred a DBN corresponding to a gene net-

work in *drosophilia melanogaster* using RJ-MCMC, with the assumption that all genes changed their connections at the same time. Of course this is not realistic, and a softer approach [157] was to infer clusters of genes with the same change points, and their network topologies in a second step.

A DBN whose topological changes were node- (gene-) specific, also corresponding to a gene network in *drosophilia melanogaster* was generated from time-series data [8] by the R-package *ARTIVA*.

ARTIVA models the network auto-regressively, where each node, apart from the first timeslice, receives general linear input from the nodes of the previous timeslice. The regression parameter values are assumed to be normally distributed, the so-called *Gaussian process prior*, with a variable covariance σ dependent only on the topology. This is not only intuitive but also renders the integration over these parameters and σ analytically tractable [8,158], so the algorithm can effectively ignore them when inferring the DBN's topology. Finding the regression parameters is optional, and *ARTIVA* derives them from topology only at the end of the algorithm for final output. This approach was developed in the context of signal processing [158], where the contributions from each frequency were taken to be normally distributed, given a set of input frequencies which were inferred by RJ-MCMC.

For our purposes the Gaussian process prior requires the introduction of latent variables, as it implicitly requires the equation parameters to also have a Gaussian distribution[2]. This was not an issue for ARTIVA's authors [8], whose input was gene-expression data with a continuous range, but the CAS cohort also includes binary and interval data. We employed the well-known approach of Albert and Chibb [9], replacing every discrete target variable Y_i with a continuous variable Z_i, where Y_i equals Z_i rounded off to the nearest integer value. Accordingly, the cutoff values for Z_i are at the half-integers. The minimum integer (usually zero) Y_{min} has a corresponding Z-range $(-\infty, Y_{min} + 0.5)$ and the maximum integer Y_{max} has a corresponding Z-range

[2]Our thanks to Prof. S. Wood, then of the Melbourne Business School, for alerting us to this issue and advising this solution.

of $(Y_{max} - 0.5, \infty)$. This is illustrated by the figures in appendix B.

At each iteration the values of Z_i are randomly selected from a truncated Gaussian distribution at each iteration, so that the MCMC algorithm is also sampling over random Gaussian distributions with preference to the more probable ones. The mean is given by the predictor variables $\mathbf{X}$ and the linear coefficients a_{ij} at that iteration, while the truncations are at the boundaries of the appropriate intervals. The complete development was originally presented in [9].

ARTIVA's output has the useful feature of providing the posterior probability for each edge it infers. While its graphs are generated by taking a 50% cutoff, we used this information to identify meaningful subsets within the data.

The underlying mathematics is simplified by assuming that the linear coefficients connecting nodes have a Gaussian distribution, otherwise known as the Gaussian process prior.

The incoming edges for each gene are modified at specific timeslices called changepoints, for which each gene has its own set. Combinations of incoming edges are inferred for each set of changepoints, where the algorithm moves through the space of possible edge combinations by either adding or removing possible edges. Likewise, sets of changepoints are sampled by RJ-MCMC, where the permitted steps are the addition, removal or shifting of a changepoint. A similar approach was taken by Andrieu and Doucet [158] to infer the number of wave inputs to a given signal.

In this work we used *ARTIVA* by replacing gene activity with data from the CAS longitudinal cohort study [5, 6], taking advantage of *ARTIVA*'s ability to infer network changes [8], as previous studies [19] indicate critical periods in asthma development. The deficiency of *ARTIVA* is its assumption of linear relationships. This was of particular concern to our work because the dynamics of asthma development contain critical thresholds at which development switches from one track to another, such as the previously mentioned example of house dust mite *IgE* titers in the second *year-of-life*. We overcame this by placing conditions on the data to distinguish between those who exhibit *wheeze* in the fifth *year-of-life* from those who do not, and thus identifying separate

paths of development corresponding to different asthma phenotypes. In section 5.7 we find that networks of *IgE* titres varied between those who developed *wheeze* in the fifth *year-of-life* and those who did not. We found a similar result for interleukin titres.

2.3 Classification and prediction

Some statistics-based algorithms called classifiers can be trained to recognise classes of objects [121, 127, 128]. A set of variables, called predictors, whose values are distributed differently between the classes, are used to make the classification. The values of these predictor variables are distributed differently between the classes, as expected in a Bayesian network, and are used to make the classification. An illustrative example of such a classifier is the spam filter, in which incoming emails are classified as being either spam or legitimate according to their observable properties. Similarly, clinicians routinely use medical tests to determine if a patient has a particular disease [131, 143].

Prediction is a special class of classification, in which the response variable is concerned with a future property. The future responses of interest to us were centred on the manifestation of *wheeze* in the fifth *year-of-life*, which we take as our proxy for asthma. We also consider more specific responses such *wheeze* accompanied by *atopy*, or *atopic-wheeze*, and *wheeze* not accompanied by *atopy*, or *nonatopic-wheeze*. All of these responses are binary, being either true or false, or positive and negative for the trait in question. We therefore restrict all discussion of prediction to binary predictors for the rest of this work.

2.3.1 Training and testing

If the predictor distributions for each class are known, then there are several methods available to find the probability of each class containing the object being classified. Usually these distributions need to be determined. The required parameters for each classifier are found from a test set, a set of samples whose predictors and class are known. The inferred distributions are then used to find the probability of the subjects belonging to each class. Classifiers are divided into two categories: [159, 160]

2.3.2 Discriminative classifiers and logistic regression

A discriminative classifier is one which specifies a cutoff and classifies according to which side of the cutoff the sample is on. In its simplest form, one uses the midpoint between the means of the two distributions. If we wish to set a cutoff with a probability other than 50%, then this may calculated either by linear regression, or by logistic regression.

Logistic regression has an S-shaped functional dependence, instead of a linear one, so the predictor value is related to the probability by $log(\frac{p}{1-p})$. This approximates the step function, which is a better choice for discriminating between two alternatives.

2.3.3 Generative classifiers and naïve Bayes

A generative classifier is one which trains by finding the class-dependent probability distributions. This can rarely be done exactly, so it is common to assume that the distributions are Gaussian. It is also common to assume that the predictors are distributed independently, or equivalently, that they do not interact. While this is frequently not the case, and has been shown to be detrimental in some problems [122, 161], it has also been found to work surprisingly well [121, 125, 126]. In fact, some studies have found that the effort of allowing for interactions can lead to a poorer performance [124].

A generative classifier with these two assumptions is called naïve Bayes.

2.3.4 Generative vs discriminant classifiers

Whether it is the generative or the discriminative classifier which performs better often depends on the data in question. For small-to-moderate numbers of predictors which approximately conform to the assumptions of the classifier, it is common for them to perform equally well. Often the assumptions are not well respected, in which case the discriminitive classifier would usually have the advantage [127]. In practice it is usually necessary to make assumptions about the distributions, such as the Gaussian assumption of naïve Bayes. This is a vulnerability not shared by discriminative classifiers.

Generative classifiers do hold an advantage when there are a large number of predictors. This is due to a common but counter-intuitive phenomenon in machine learning

called *overfitting*, in which adding additional predictors degrades the classifier's performance. This can be understood as the classifier fitting the data too well, so that it matches the random fluctuations in the data which are really just noise. Discriminative classifiers are quite vulnerable to overfitting, especially with small sample sizes or poor predictors [162]. Generative classifiers are theoretically immune to it [113], but can still be affected if the data is poor or does not match the classifier's assumptions very well.

The issue of which classifier is better, and under what circumstances is still controversial [159, 160, 163].

2.3.5 Assumption of independence

It is common to use linear classifiers which assume that the predictors are completely independent with no correlation between them. While greatly simplifying the algorithm, neglecting these interactions comes at the cost of ignoring potentially important information. This is especially true for Bayesian classifiers which require the entire distribution in principle. While this cost is not always small [117, 131], the literature abounds (*e.g.* [123, 127, 128]) in applications in which naïve Bayesian classification works considerably better than one might intuitively expect. This was eventually understood by an information theoretic analysis [126], which found the loss of information to be the most relevant factor. Other authors argued that attempts to account for correlation were rarely worth the cost [125], and sometimes self-defeating [124]. The classifiers used in this work are linear and also make this assumption.

2.3.6 Measuring classifier performance

Classifier performance can be measured in a number of different ways, based on the fraction of correctly identified cases and/or controls. Which is most appropriate often depends on the question at hand, and the reader is referred to [162] and references therein for a thorough discussion. Some commonly used measures are:

True positive rate: the fraction of positive samples which are found to be positive,

True negative rate: the fraction of negative samples which are found to be negative,

False positive rate: the fraction of negative samples which are found to be positive,

False negative rate: the fraction of positive samples which are found to be negative,

Positive predictive rate: the fraction of positive findings which are truly positive,

Negative predictive rate: the fraction of negative findings which are truly negative.

ROC plot and AUC

While the logical cutoff for differentiating between case and control is at a probability of 0.5, there are several reasons for choosing differently. For example, a lower cutoff might be chosen if the cost of a false negative is high. An obvious example of this is a screening test for a serious medical condition. If the condition is missed then the patient's health and even life are at risk so the clinician chooses to err on the side of caution until a more definitive test is performed. Similarly, a high cutoff might be used if there is a significant cost associated with false positives and standards of evidence are required to be high.

One may find the true and false positive rates as a function of the cutoff and plot it on a pair of axes, as shown in figure 2.1. Such a plot is called a ROC plot, and the corresponding curve a ROC curve. ROC stands for Receiving Operator Characteristic, dating to its origins as a way of classifying the accuracy of radar operators in the second world war.

The area under the ROC-curve, or AUC, is another measure of classifier performance, and may be interpreted as the probability that the case will be ranked above the control when presented with a case-control pair [164].

Cross-validation

Ideally one trains the classifier on one dataset, called the "training set", and then tests it on an independent dataset, called the "testing set". If only one dataset is available, as in our work, then we may divide it into training and testing sets. (See [165] and references therein for history and a formal description.) Since limited training data is a limiting factor on classifier performance, the division is rarely even. Rather, the training set

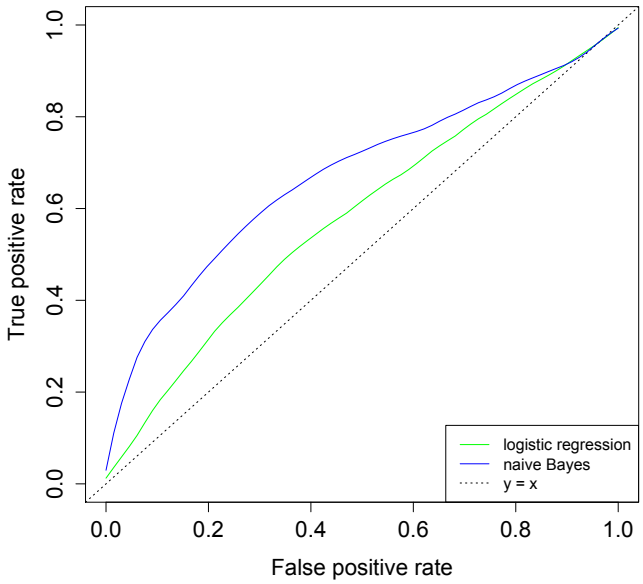


Figure 2.1: **ROC plot showing ROC curves for logistic regression and naïve Bayes classifiers predicting *wheeze* in the fifth *year-of-life*:** The predictors are a randomly chosen assortment from CAS.

is typically chosen to be from two to ten times the size of the testing set, depending on the details of the problem. While the division is random, it is also a good idea to maintain the overall ratio of cases to controls in the two subsets. We found the best compromise between training and testing with a two-to-one ratio between training and testing sets so that the testing set was one third of the dataset, but other values are possible (*e.g.* [166]).

This works well but is vulnerable to statistical fluctuations among the possible partitions. The natural remedy is to cycle through all possible partitions, or at least enough of them to smooth out any statistical outliers. A common practice, which we also employ, is to perform the described procedure 100 times, choosing a different partition each time, and average over the results.

Other solutions to this problem are in common use, such as "one with replacement", where one sample is removed for testing and replaced for the next iteration. We have mentioned these other methods for the sake of discussion but they have not been em-

ployed in our work. A review is available in [166].

Summary

We have just described the network-based methods with which we shall study the data-intensive CAS cohort. The first was a description of *ARTIVA*, which we have adapted to handle count data such as the numbers of different types of infection. Next, we described basic classifiers, based on a very simple BN, their application to prediction, and the methods of measuring and minimising their errors. We apply these methods in the next chapter to not only identify which variables make good predictors, but to probe the endotypic structure of asthma by identifying those which exclusively predict specific types.

Chapter 3

Asthma endotypes from early-life prediction

3.1 Introduction

As discussed in section 2.1, asthma comes in a variety of forms or *endotypes* [167, 168], whose phenotypic variation can range over characteristics as varied as triggers, severity, medication responses, as well as various immunity-related factors. It is not uncommon to see discussion of eosinophilic *vs* neutrophilic hypersensitivity [3, 15, 169], and there have been increasing calls in the literature [168–170] to consider the more specific endotypes of asthma as characterised by accompanying comorbidities.

We begin by demonstrating that the *atopic* and *non-atopic wheeze* endotypes (remember fifth-year *wheeze* is our proxy for asthma) have different predictors, in agreement with [22], and find optimal predictor sets for each. (Unless stated otherwise, all predictors were taken from the first two *years-of-life*.) The best predictor set for fifth-year *atopic-wheeze* was also the best one for *wheeze* in general. This was not surprising as one of the best known predictors of asthma is derived from titres of *house-dust-mite IgE* in the second *year-of-life*. Also, CAS participants were chosen for a family history of *atopy* or asthmatic *wheeze*. Furthermore, the prediction by (log of) *house-dust-mite IgE* of fifth-year *atopic-wheeze* was better, as measured by the AUC, than its prediction of fifth-year *wheeze*. We therefore considered that *house-dust-mite IgE*'s predictive ability for *wheeze* is really limited to *atopic-wheeze*, so that prediction of *wheeze* in general would be less effective in a different data set with a lower proportional of *atopic*s among those with fifth-year *wheeze*.

To demonstrate this we derive equations relating the AUCs from the same clas-

sifier/predictor combination predicting a general case, say *wheeze*, and a specific endotype, say *atopic-wheeze*. We then go on to construct an index of this exclusivity and used it to identify a new endotype of *atopic-wheeze*, and exclusive predictors of *nonatopic-wheeze*.

3.2 Identification of *wheeze* endotypes in the fifth *year-of-life* through exclusive early-life predictors

We began by seeking the best set of predictors that we could find for *wheeze* in the fifth *year-of-life*. In accordance with the findings of appendix A, we chose our variables in order of decreasing predictive ability unless the next predictive variable was expected to be correlated with one that had already been chosen, a compromised greedy algorithm. This yielded respectable predictions of fifth-year *wheeze* and better predictions for *atopic-wheeze*, as measured by the AUC. Repeating this process for fifth-year *nonatopic-wheeze* also allowed us to make good predictions.

3.2.1 Optimal predictor set for fifth-year wheeze

Plotting AUC as a function of using the top one-to-twelve predictors for fifth-year *wheeze* for both logistic regression and naïve Bayes (figure 3.1), we found a best AUC of .77 from the top three predictors. (All in the second *year-of-life*, these were (log of) *house-dust-mite IgE*, (log of) *infant-phadiatop-IgE*, and *severe-LRI*.) Attempts to include more than these three predictors led, apart from a slight bump at five predictors, to a steadily lower performance. This diminution in performance from adding additional predictors was in qualitative accordance with the simulations in appendix A.

Now *house-dust-mite* is an airborne allergen. If its predictive ability is limited to *atopic* endotypes, then the AUC is going to be higher for those specific cases and .5 for the remaining cases.

We explore this in a more general and quantitative sense in what follows and find that we can use it to identify biologically distinct endotypes, each with its own specific predictors, causes, and aetiology.

3.2.2 Predictor set for fifth-year atopic-wheeze

The AUC scored by (log of) *house-dust-mite IgE* in the second *year-of-life* when predicting fifth-year *atopic-wheeze* was .79, while the top five predictors of fifth-year *wheeze*, namely (log of) *house-dust-mite IgE*, (log of) *infant-phadiatop-IgE*, *severe-LRI*, *febrile-LRI* and *severe-viral-LRI*, all from the second *year-of-life*, combined to score an AUC of .83 when predicting *atopic-wheeze* in the fifth *year-of-life*.

Counting early-life *atopies* to predict fifth-year *atopic-wheeze*

The top twelve predictors of *atopic-wheeze* were *IgE* titres (or their logarithms) for *house-dust-mite*, *peanut*, or *infant-phadiatop*, but combining them simply did not lead to

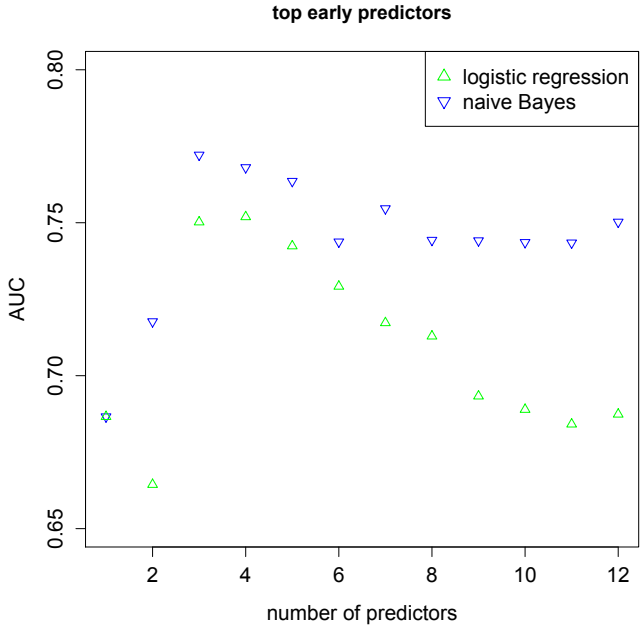


Figure 3.1: **Effect on logistic regression and naïve Bayes AUC as successive CAS predictors are added to the set of predictors:** The predictors were added in order of decreasing AUC. The performance peaked at three predictors and fell away more quickly for logistic regression, as expected from our discussion on overfitting in section 2.3 (see also appendix A). The top 12 predictors, all of which were taken from the second *year-of-life* except where otherwise indicated, were, in decreasing order of predictive power, (log of) *house-dust-mite IgE*, (log of) *infant-phadiatop-IgE*, *severe-LRI*, *febrile-LRI*, *severe-viral-LRI*, (log of) *peanut-IgE* (first *year-of-life*), *febrile-viral-LRI*, (log of) *peanut-IgE*, *LRI*, *infant-phadiatop-IgE*, *wheeze*, and *airborne-atopy*.

an improved AUC. Even combining the top *peanut-* and *house-dust-mite- IgE* predictors led to no improvement, which suggested to us that either they are each carrying near-redundant information about the response, or that their class-dependent distributions are sufficiently correlated to generate errors in the linear classifiers. The correlations among the cases and the controls were .77 and .35, respectively. While these correlations are indeed significant, they are much moreso among cases. This difference suggests that the true risk factor was that both predictors were elevated, given that higher *IgE* titres correspond to *atopic-wheeze* risk [6].

We therefore considered whether the number of allergens which elicit an atopic response, or *atopy-number*, might be a predictor of *atopic-wheeze* in the fifth *year-of-life*. (Note that *infant-phadiatop* and *phadiatop* are not included in this definition to avoid double-counting.) Interestingly, *atopy-number* from the first *year-of-life* was a better predictor than from the second. This single predictor, in the first *year-of-life*, scored an AUC of .81 predicting fifth-year *atopic-wheeze* and an AUC of .66 predicting fifth-year *wheeze*, second only to the .67 scored by second-year (log of) *house-dust-mite IgE*.

We also considered the number of airborne allergens to which the participants were *atopic*, or *aeroatopy-number*. Its predictive power was greater in the second *year-of-life* (AUC .74) than in the first (AUC .65), but was still lower than for first-year *atopy-number*.

We note in passing that when we combined first-year *atopy-number* with *febrile-LRI* from the second *year-of-life* to predict fifth-year *atopic-wheeze* the resulting AUC was .86, well into the excellent range.

3.2.3 Optimal predictor set for fifth-year nonatopic-wheeze

The best predictors for *nonatopic-wheeze* in the fifth *year-of-life* were, in decreasing order of AUC from logistic regression; *number-of-older-children-in-the-home-at-birth* (.68), (log of) first- and second- year *infant-phadiatop-IgE* (.65), then *number-of-older-children-in-the-home* (.64), number of *HRV* (*human rhinovirus*) infections and *trecs* (*T-cell receptor excision circles*) (.62) from the second *year-of-life* (.63). We mention

in passing that the optimal predictor set for *nonatopic-wheeze* in the fifth *year-of-life* was *number-of-older-children-in-the-home-at-birth*, (log of) *infant-phadiatop-IgE* from the first *year-of-life*, then (log of) *infant-phadiatop-IgE*, *HRV* and *trecs* from the second *year-of-life*, with an AUC from naïve Bayes of .75.

3.3 Construction of exclusivity index

Asthma endotypes may be differentiated by the demonstration of separate predictors. We now proceed to construct and explore an index whose function is to indicate whether an apparent predictor for a general condition, such as *wheeze*, is really only predictive of a given endotype of that condition, such as *atopic-* or *non-atopic- wheeze*. While identifying useful predictors is an obvious incentive for such a tool, we consider the real value to be in identifying specifically predictable endotypes, reasoning that if it does have a specific predictor then it is likely to be of biological and/or clinical significance.

3.3.1 Relations between AUCs of general and specific conditions

The AUC is a measure of the accuracy of a classifier, in combination with the predictors being fed into it. For a given classifier, such as the logistic regression and naïve Bayes classifiers that we have chosen, it is a measure of the quality of the predictor(s) being used. As explained in subsection 2.3.6, the AUC is defined as the area under the ROC curve, but it has another interpretation [164]. If a trained classifier/predictors combination is presented with a pair of samples, in which one is a case (positive response), and the other a control (negative response), then the AUC gives the probability of the classifier/predictors combination finding a higher probability for the case than for the control. In simpler terms, the classifier's probability of correctly sorting case from control by simply playing the odds.

Let us consider a sample set whose cases may be neatly divided into two or more sub-cases $a, b, \ldots$. For concreteness, the general case could be fifth-year *wheeze* divided into fifth-year *atopic-wheeze* (subcase a) and *nonatopic-wheeze* (subcase b). For a given classifier, suppose P_1 scored an AUC of auc_{1a} predicting sub-case a but could not distinguish other general cases from controls. The case-control pairs can then be divided

into those whose case is a member of a, and those whose is not. The former is predicted with an AUC of auc_{1a}, the latter with an AUC of .5 which is equivalent to random guessing. So if P_1 is used to predict the general case, then it would score an AUC of

$$AUC_1 = q_a \, auc_{1a} + (1 - q_a) \times .5, \tag{3.1}$$

where q_a is the proportion of cases belonging to sub-case a.

Similarly, if there is a second sub-case b, which does not intersect with a, for which predictor P_2 scores an AUC of auc_{2b} while unable to distinguish other cases, including a, from controls, then the AUC of P_2 predicting the general case would be

$$AUC_2 = q_b \, auc_{2b} + (1 - q_b) \times .5, \tag{3.2}$$

where q_b is the proportion of cases belonging to sub-case b.

If P_1, P_2 are taken together to predict the general case, then the anticipated AUC obeys

$$\begin{aligned} AUC_{12} &= q_a \, auc_{1a} + q_b \, auc_{2b} + (1 - q_a - q_b) \times .5 \\ &= AUC_1 + AUC_2 - .5 \end{aligned} \tag{3.3}$$

since, in the ideal scenario, q_a of the cases are distinguished with an accuracy of auc_{1a}, q_b with an accuracy of auc_{2b}, while the rest are randomly guessed. If either of P_1, P_2 are predictive of each others' subclass then that will increase AUC_{12}, but that scenario is forbidden by equations (3.1) and (3.2). AUC_{12} is degraded by the introduction of noise with each additional predictor. The noise is due to P_1 randomly guessing cases of b and P_2 randomly guessing cases of a. Further analysis is beyond the scope of our requirements.

Necessary and sufficient conditions for the exclusivity index to hold

The above derivation of equations (3.1) to (3.3) begins with the assumption of exclusivity, and as such it only demonstrates their necessity for the predictor to be exclusive. In practice, calculations of AUC vary slightly from one calculation to another, which we address below by calculating error margins based on AUC uncertainties of $\pm.01$. There is, in practice, another factor affecting these equations. If the number of sub-cases is too small a fraction of the more general case, then prediction of the general case can be adversely affected by noise.

If the predictor is not exclusive to the subcase a then equation (3.1) will underestimate the predictor's score with the general case AUC_1. One might therefore expect equation (3.1) to be a lower bound. The effect on equation (3.3) is more complicated. If, contrary to founding assumptions, there is overlap between cases a and b then the effect on AUC_{12} is harder to ascertain. On the one hand, there is less new information being brought to the prediction leading to a lower general AUC. On the other hand, where cases a and b overlap the predictors might combine constructively, so that classification of either case is more accurate with both predictors. The combination might also be destructive.

The overall picture then is that any deviation from the founding assumptions used to derive equations (3.1) to (3.3) will introduce errors, but these might be cancelled by noise. We therefore take these equations, and the indices we now derive from them, as strong indicators that the predictor in question is exclusively predictive of the endotype in question. Care must be taken in the case of rare endotypes whose low case numbers render them vulnerable to statistical noise. We dealt with this by only using AUCs found to be statistically significant by the Mann-Whitney U-test. The corresponding p-values are shown for the rare endotypes we discuss in table 3.5.

3.3.2 Relations among specific predictors

We can also approximate the accuracy with which P_1 would predict cases of b, and P_2 cases of a. Since the sub-cases are distinct, the sub-case a also forms a subset of

non-cases of b. No case of a is also a case of b, so when considering the performance of P_1 predicting b, it is better to break the non-cases of b into a and non-a, predicted with accuracy auc_{1a} and 0.5, respectively. Recalling that P_1 will distinguish cases of a from cases of b, auc_{1a} of the time, and be randomly guessing otherwise, the AUC of P_1 predicting b will be

$$\begin{aligned} auc_{1b} &= q_a auc_{1a} + (1 - q_a - q_b) \times .5 \\ &= AUC_1 - q_b \times .5, \end{aligned} \tag{3.4}$$

where the final line substitutes equation (3.1). There is a corresponding result predicting a with predictor P_2.

We illustrate with the example of predicting *nonatopic-wheeze* in the fifth *year-of-life* from (log of) *house-dust-mite IgE* in the second *year-of-life*. We shall show in section 3.4 that this predictor is exclusive to a specific endotype of *airborne-atopy*, with absolutely no overlap with *nonatopic-wheeze*. Substituting the relevant values into equation (3.4) finds that this prediction should score an AUC of .47. (The actual AUCs were .49 (logistic regression) and .54 (naïve Bayes). Since these two classifiers usually score almost identically for single predictors, and given the irregular shape of the ROC curve (not shown), we attributed the discrepency to a breakdown in the linear relationship between the predictor and the outcome.) This suggests that a predictor which scored an AUC below 0.5 for a particular response might be predictive of a separate response. However, there is no clear indicator of what that response would be, and since there appears to be no predictive value in this approach we do not pursue it any further.

3.3.3 Index of predictor exclusivity

Our purpose here is to identify distinct endotypes along with their exclusive predictors. We are not primarily concerned with the value of the AUC, apart from using them in our calculations, so we have re-arranged equations (3.1-3.4) to put factor .5 on the right-hand-side. The left-hand-side then becomes a measure of how exclusive the predictor is to the endotype in question, so we have named it the "exclusivity index", being .5

general case		*wheeze*		*atopic-wheeze*
endotype		*atopic-wheeze*	*aeroatopic-wheeze*	
variable	year	(±.05)	(±.04)	(±.13)
(log of) *house-dust-mite IgE*	1	**.45**	**.48**	**.60**
(log of) *house-dust-mite IgE*	2	**.49**	**.51**	**.58**
(log of) *infant-phadiatop-IgE*	2	.57	.62	.84
peanut-IgE	1	**.51**	.57	.81
atopy-number	1	.41	**.50**	.81
atopy-number	2	**.51**	**.52**	**.57**
severe-LRI	2	.71	.74	.87
febrile-LRI	2	.64	.71	1

Table 3.1: **Results from substituting AUCs from some of the better predictors into the exclusivity index (equation (3.5))**. For cases of fifth-year *wheeze* we have considered both *atopic-wheeze* and *aeroatopic-wheeze*. We have also considered cases of *atopic-wheeze* with the endotype *aeroatopic-wheeze*. The error margins given at the top of the data columns were calculated by substituting an error margin of ±.01 into both AUCs in equation 3.5, and rounding up to the nearest .01. (Log of) *house-dust-mite IgE* in the first two *years-of-life* apparently predicts *aeroatopic-wheeze* exclusively, while *peanut-IgE* in the first *year-of-life* appears to predict *atopic-wheeze*, unsurprisingly, but not *aeroatopic-wheeze*. The .50 scored by *atopy-number* in the first *year-of-life* concerning the *aeroatopic* subset of *wheeze* is probably a fluke, since it scores badly for the larger *atopic* subset.

when the predictor is exclusive to within a predetermined error margin. That margin was found by allowing both AUC_1 and auc_{1a} to vary by ±.01.

So, if predictor P_1 is exclusively predictive of endotype a then re-arranging equation (3.1) gives

$$\frac{AUC_1 - q_a\, auc_{1a}}{1 - q_a} = .5 \tag{3.5}$$

with an error margin of $\pm.01 \times \frac{1+q_a}{1-q_a}$.

When considering the combination of two non-intersecting endotypes, rearranging equation (3.3) gives

$$AUC_1 + AUC_2 - AUC_{12} = .5 \pm .03. \tag{3.6}$$

variable	year	index
number-of-older-children-in-the-home-at-birth	birth	**.47**
(log of) *infant-phadiatop-IgE*	1	.59
(log of) *infant-phadiatop-IgE*	2	.73
number-of-older-children-in-the-home	1	**.49**
number-of-older-children-in-the-home	2	**.51**
HRV	2	**.47**
severe-LRI	2	.68
febrile-LRI	2	.69

Table 3.2: **Results from substituting AUCs from some of the better predictors of *nonatopic-wheeze* into equation (3.5).** We are considering cases of fifth-year *wheeze* and the endotype of *nonatopic-wheeze*. An error margin of ±1 has been allowed for each AUC contributing to the index calculation, leaving a final margin of ±.03.

	atopy-number	*peanut-IgE*
number-of-older-children-in-the-home (birth)	.46	**.52**
number-of-older-children-in-the-home (year 2)	**.50**	.54
HRV (year 2)	**.50**	.60

Table 3.3: **Index scores for combining predictors of *atopic-wheeze* from the first *year-of-life* and predictors of *nonatopic-wheeze* to predict *wheeze*.** An error margin of ±1 has been allowed for each AUC contributing to the index calculation, leaving a final margin of ±.03. As in table 3.1, *atopy-number* and *peanut-IgE* from the first *year-of-life* appears to be less-than-solid as exclusive *atopic-wheeze* predictor. The other variable pairs conform to equation (3.6).

	house-dust-mite IgE (log)	*atopy-number*	*severe-LRI*	*febrile-LRI*
number-of-older-children-in-the-home (birth)	**.48**	**.49**	.55	**.51**
number-of-older-children-in-the-home (year 2)	**.50**	**.50**	.56	**.53**
HRV (year 2)	**.50**	**.52**	.55	**.52**

Table 3.4: **Index scores for combining predictors of *atopic-wheeze* from the first *year-of-life* and predictors of *nonatopic-wheeze* to predict *wheeze*.** An error margin of ±1 has been allowed for each AUC contributing to the index calculation, leaving a final margin of ±.03. Consistent with table 3.1, the number of neither *febrile-* nor *severe- LRI*s predicts either endotype exclusively. The other variable pairs conform to equation (3.6).

3.3.4 Illustration of exclusivity equations with atopic- *and* non-atopic-wheeze *endotypes*

We have seen that the best predictors of *wheeze*, namely the *IgE* titres, are clearly of relevance to *atopic-wheeze*, with the corresponding AUCs higher for the *atopic-wheeze* condition than for the *wheeze* condition. Equation (3.1) allows us to investigate whether their predictive ability is limited to *atopic-wheeze*, so that they are unable to distinguish *nonatopic-wheeze* from *non-wheeze*. Equivalently, do the predictors in question derive their efficacy for the general case from their predictive *atopic* endotype being a majority subset (of the 56 participants with fifth-year *wheeze*, 35 were atopic and 30 were aeroatopic)?

For reasons that will become clear, we use *house-dust-mite IgE* in the second *year-of-life* as our illustrative example. Its relevance to *aeroatopic-wheeze* should surprise noone, but other aeroallergens do not serve this purpose nearly as well. We also point out that this is more than predicting later *atopy* from current *atopy*, as there is no reason for thresholds related to fifth-year outcomes to match the thresholds for early-life conditions.

Applying equation (3.1), (log of) *house-dust-mite IgE* from the second *year-of-life* scored an AUC of .76 for prediction of fifth-year *atopic-wheeze*. For predictions of the more general fifth-year *wheeze* the above argument predicts an AUC of

$$.76\frac{35}{56} + .50\frac{21}{56} = .66 \approx .67, \tag{3.7}$$

in agreement with explicit calculation. Similarly, (log of) *house-dust-mite IgE* from the second *year-of-life* scored a higher AUC of .79 predicting fifth-year *aeroatopic-wheeze*, a subset of fifth-year *atopic-wheeze* with 30 cases. Equation (3.1) again holds, namely

$$.79\frac{30}{56} + .50\frac{26}{56} = .66 \approx .67. \tag{3.8}$$

One might also consider *aeroatopic-wheeze* as a subset of *atopic-wheeze*, yielding

$$.79\frac{30}{35} + .5\frac{5}{35} = .75 \approx .76, \tag{3.9}$$

where the difference is well-within natural variation of AUC and round-off errors are worse due to the smaller number of cases.

In such a scenario we can derive an upper bound on how well the general response can be predicted from variables which predict subsets. Given a perfect predictor of *aeroatopic-wheeze* (AUC=1), its AUC for predicting fifth-year *wheeze* is then

$$\frac{30}{56} + .5\frac{26}{56} = .768. \tag{3.10}$$

The corresponding result from an ideal predictor of fifth-year *atopic-wheeze* predicting fifth-year *wheeze* is

$$\frac{35}{56} + .5\frac{21}{56} = .813. \tag{3.11}$$

A more realistic but still optimistic predictor of *aeroatopic-wheeze*, with an AUC of say, .9, would score an AUC of

$$.9 \times \frac{30}{56} + .5\frac{26}{56} = .714, \tag{3.12}$$

predicting fifth-year *wheeze*. The corresponding figure for *atopic-wheeze* is .75.

To identify valid endotypes we require not just that they be defined by biological indicators, but that they have an exclusive predictor which is unpredictive of the non-endotype. We illustrate this by using (log of) *house-dust-mite IgE* from the second *year-of-life* to demonstrate that *atopic-wheeze* and *aeroatopic-wheeze* from the fifth *year-of-life* are well-defined endotypes in their own right.

Substituting in the AUCs scored by second-year (log of) *house-dust-mite IgE* for

fifth-year *wheeze* and *atopic-wheeze* into equation (3.5) gave us

$$\left(.67 - .76 \times \frac{35}{56}\right) \times \frac{56}{21} = .49 \pm .05. \tag{3.13}$$

confirming that *atopic-wheeze* is a well-defined endotype of *wheeze*. The result is similar for *aeroatopic-wheeze* as an endotype of *wheeze*, with

$$\left(.67 - .79 \times \frac{30}{56}\right) \times \frac{56}{26} = .48 \pm .04. \tag{3.14}$$

Consistency requires that *aeroatopic-wheeze* be an endotype of *atopic-wheeze*. Substituting in the appropriate AUCs gave us

$$\left(.76 - .79 \times \frac{30}{35}\right) \times \frac{35}{5} = .60 \pm .13, \tag{3.15}$$

as required, although the errors are substantially larger in this example.

In subsection 3.2.2 we found that *atopy-number* in the first *year-of-life* scored AUCs of .66 and .80 predicting fifth-year *wheeze* and *aeroatopic-wheeze*, respectively. Substituting into equation (3.1) finds

$$.80\frac{30}{56} + .50\frac{26}{56} = .67. \tag{3.16}$$

The corresponding equation for the *atopic-wheeze* endotype does not hold, with first-year *atopy-number* predicting fifth-year *atopic-wheeze* scoring the same AUC as *aeroatopic-wheeze* despite a different number of cases. *Atopy-number* from the second *year-of-life*, by contrast, obeys equation (3.1) for all combinations of *wheeze*, *atopic-wheeze* and *aeroatopic-wheeze*, as indicated in table 3.1. This was unexpected as the first-year predictor was the better predictor of *atopic-wheeze*.

We have calculated the exclusivity to *atopic-wheeze* and *aeroatopic-wheeze* of several good predictors, and presented them in table 3.1. Not surprisingly, (log of) *house-dust-mite IgE* is also an exclusive predictor in the first *year-of-life*. Another exclusive

response atopy	predictor allergens	AUC	p-value	nb. cases
cat	(log) dust-mite	.94	$\mathbf{2.2 \times 10^{-7}}$	9
peanut	(log) dust-mite	.96	$\mathbf{9.7 \times 10^{-9}}$	10
couch	(log) dust-mite	.97	$\mathbf{1.3 \times 10^{-11}}$	14
rye	(log) dust-mite	.93	$\mathbf{1.7 \times 10^{-11}}$	16
cat/peanut/couch/rye	(log) dust-mite	.94	$\mathbf{7.8 \times 10^{-14}}$	20
mould	mould	.78	$\mathbf{1.2 \times 10^{-6}}$	7
mould	mould (year 3)	.90	$\mathbf{2.4 \times 10^{-9}}$	7
mould & dust-mite	(log) mould and dust-mite	.94	$< \mathbf{2.4 \times 10^{-9}}$	5
dust-mite	(log) dust-mite	.81	$\mathbf{3.8 \times 10^{-10}}$	28
dust-mite	(log) dust-mite, *febrile-LRI*	.85	$< \mathbf{3.8 \times 10^{-10}}$	28

Table 3.5: **Predictors of *atopic-wheeze* endotypes from *IgE* data:** Predictors taken from second year of life unless indicated otherwise. p-value is found from the Mann-Whitney U-test for single predictors.

response atopy	*wheeze*	*atopic-wheeze*	*aero atopic-wheeze*
cat	.62 ±.02	.70 ±.02	.73 ±.02
peanut	.61 ±.02	.68 ±.02	.71 ±.02
couch	.57 ±.02	.62 ±.03	.63 ±.03
rye	.57 ±.02	.62 ±.03	.63 ±.04
cat/peanut/couch/rye	**.52** ±.03	**.52** ±.04	**.49** ±.05
house-dust-mite	**.53** ±.03	**.56** ±.09	**.51** ±.29
mould (second *year-of-life mould-IgE*)	.46 ±.02	**.48** ±.02	.44 ±.02
mould (third *year-of-life mould-IgE*)	**.51** ±.02	**.52** ±.02	.55 ±.02

Table 3.6: **Exclusivity index of (log of) *house-dust-mite IgE* in the second *year-of-life* for various allergen-specific *atopic-wheeze*s.**

predictor is *atopy-number*, though only in the second *year-of-life*. *Febrile-* and *severe-LRI*s are not exclusive to *atopic-wheeze* despite their efficacy in predicting it.

3.4 New asthma endotypes, characterised by their *atopic* triggers

The CAS data set includes neither neutrophilic nor eosinophilic data, but we obtained very encouraging predictor scores by considering *atopy* to specific allergens. *Cat*, *peanut*, *couch* and *rye* were particularly well-predicted, with AUCs greater than .9. As a guard against the small number of cases we also found the significance of each AUC using the Mann-Whitney U-test. These p-values are included in table 3.5 and range from 1.2×10^{-6} to 7.8×10^{-14}.

Calculating the exclusivity index from equation (3.5) for each of these allergen-

specific endotypes found three smaller endotypes with exclusive predictors, as shown in table 3.6. These were *wheeze* with *mould*-related *atopy*, *wheeze* with *house-dust-mite* related *atopy*, and *atopic-wheeze* where the *atopic* allergens include at least one of *cat*, *peanut*, *couch* or *rye*. (We observed that participants with *atopic-wheeze* sensitized to one or more of these allergens in the fifth *year-of-life* were also *atopic* to *house-dust-mite*.) We call this last case *multi-allergen atopic-wheeze*. The first was exclusively predicted as an endotype of both *wheeze* and *atopic-wheeze* by *mould-IgE* in the third *year-of-life*. However it was not found to be an exclusive endotype of *aeroatopic-wheeze*, which it logically should be given that *mould* is an airborne allergen. Plausible explanations for this include statistical noise due to the small number of cases, and cases of *mould*-related *atopic-wheeze* strongly coinciding with some more relevant factor. Hence some uncertainty remains concerning the nature of this endotype.

The other two of the smaller endotypes above were exclusively predicted by (log of) *house-dust-mite IgE* in the second *year-of-life*. Furthermore, multi-allergen *atopic-wheeze* as an endotype of *house-dust-mite atopic-wheeze* scored an exclusivity index of .485.

So not only is predictive ability of *house-dust-mite IgE* limited to *aeroatopic-wheeze*, but to the even more specific endotype of *multi-allergen atopic-wheeze*. In particular, it has no predictive power for *aeroatopic-wheeze* where the allergens do not include at least one of *rye*, *couch*, *cat* or *peanut*.

3.5 Outcomes

While the ability to predict *wheeze*, our proxy for asthma, in the fifth *year-of-life* from data from the first two *years-of-life*, was strongly limited by overfitting, it was readily apparent that even linear classifiers with single or few predictors could further illuminate the nature of asthma. The literature has begun to acknowledge [15, 168–170] the importance of asthma endotypes, and their having different predictors is no real surprise. We only considered conditions which could be readily described in terms of CAS variables, such as the presence or absence of *atopy* or specific *atopic* sensitivities. Since

such definitions are trivial to compose we required the existence of at least one exclusive predictor before claiming that such a definition had any biological meaning. We were able to identify endotypically exclusive predictors with quantitative relationships to more general cases. One of the best-known predictors of asthma, *house-dust-mite IgE* in the second *year-of-life*, was found to be truly predictive of only a specific endotype of *atopic-wheeze*. This *multi-allergen* endotype is one of the most common, and the apparent predictivity of second year *house-dust-mite IgE* of *atopic-wheeze* and *wheeze* in general was attributable to its proportion of cases in *atopic-wheeze*, and of *atopic-wheeze* in the *wheeze* cases of CAS. We did attempt a similar study of *atopic-wheeze* where *house-dust-mite* was the only *atopic* allergen, but were stopped by a lack of statistical power.

Seen in this light, exclusive predictors of other endotypes, such as *non-atopic wheeze*, are less predictive for *wheeze* in general, but are equally important in understanding the bigger asthma picture. We therefore needed to consider the exclusivity of such predictors and their predictive power for the relevant endotype, without being distracted by the possibly low AUC for *wheeze* in general. This was facilitated by the exclusivity indices in equations (3.5, 3.6), which are, to within a calculable error margin, equal to one half when the endotype is exclusively predicted within the more general case. Unfortunately this does not solve the problems associated with insufficient sample size, its best contribution to that particular problem being a stronger, quantitative, argument for future detailed studies.

This *multi-allergen atopic-wheeze* will feature in the next chapter in our discussion of the infant NP microbiome. For now we conclude this chapter with a list of its specific important outcomes:

1. That the number of allergens to which the child is *atopic* in the first *year-of-life* is an excellent predictor of *atopic-wheeze*, better even than second-year (log of) *house-dust-mite IgE*, neither of which have any predictive power for any form of *nonatopic-wheeze*.

2. Equations (3.1) to (3.6) relate the AUCs scored by a predictor for the general case and for one of its specific endotypes. Equations (3.5) and (3.6) in particular give a measure of exclusivity which does not reflect the AUC itself.

3. That the predictive ability of second-year (log of) *house-dust-mite IgE* for *wheeze* in the fifth *year-of-life* is exclusive to not just *atopic-wheeze* or even *aeroatopic-wheeze*, but to *atopic-wheeze* with at least one of *peanut*, *cat*, *couch* and *rye* included among the *atopic* allergens. (Such *atopic*s in the fifth *year-of-life* always include *house-dust-mite* among their *atopic* allergens as well.) This endotype contained no exclusive smaller one that we could find.

4. That the number of older children in the house, both at birth and in the second *year-of-life*, were exclusive predictors of *nonatopic-wheeze* in the fifth *year-of-life*.

5. That some important predictors, including *severe-LRI* and *febrile-LRI*, are not exclusive to either *atopic-* or *non-atopic- wheeze* but are predictive of both.

6. That the AUCs of predictors exclusive to separate endotypes simply add (minus .5) when they are combined in a set predicting the common general condition (equation (3.3)). The obvious examples of this are predictors sets of *atopic-* and *non-atopic- wheeze* combining to predict *wheeze*, examples of which are given in tables 3.3 and 3.4.

7. That *mould-IgE* from the third *year-of-life* exclusively predicted (AUC of .90) *wheeze* with *atopy* to *mould* in the fifth *year-of-life*, as endotypes of both fifth-year *wheeze* and fifth-year *atopic-wheeze*, but not of *aeroatopic-wheeze*. We cannot determine if this last anomoly is due to statistical noise (there were only seven cases of this endotype) or the importance of a factor closely correlated with *atopy* to *mould*.

Chapter 4

The early nasopharyngeal microbiome and later (*atopic*) *wheeze* status

4.1 Introduction

There has long been a strong association between asthma development and respiratory, especially *viral*, infections during infancy [6,48,51,52], and mounting evidence [7,86,171] for an association with the *nasopharyngeal* (NP) microbiome in early infancy. The CAS cohort therefore included detailed microbiome and virus data from NP microbiome samples taken at various time-points [5, 6, 19]. The *genera* typically found in the NP microbiome are *Alloiococcus*, *Streptococcus*, *Staphylococcus*, *Haemophilus*, *Moraxella* and *Corynebacterium* [7].

We sought to go beyond conventional linear analysis, with Teo *et al.* [7] having already found correlation between *Streptococcus* and fifth-year *wheeze*. We were also interested in whether inter-*genus* interaction might be of comparable importance to any specific *genera*. We addressed these questions with dual qq-plots, in which we compared qq-plots generated for case and control samples on the same axes. The signals we observed required no such interaction, but we did observe some *genera* to be of relevance to later *wheeze* and *atopy* under certain circumstances.

The work of Teo *et al.* [7] found that samples taken during the course of an acute respiratory infection, or *ARI* samples, have different associations from those taken in the absence of a respiratory infection, or *non-ARI* samples. We therefore considered *ARI* and *non-ARI* samples separately. We would have liked to restrict our analyses to the first seven *weeks-of-life*, but limited statistical power meant that we could only do this for *non-ARI* samples so we were forced to consider *ARI* samples from the

first 90 *days-of-life*. We found that the relative abundances of some *genera* in the NP microbiome from these data sets were predictive of some fifth-year outcomes.

Subsection 4.2.1 in particular demonstrates that *Haemophilus*, *Moraxella* and *Staphylococcus* relative abundances during respiratory infections in the first 90 *days-of-life* affect later *wheeze* and *atopy* status. In section 4.2.2 we confirm the importance of *Streptococcus* to fifth-year *wheeze* status [7].

Section 3.4 identified the *multi-allergen atopic-wheeze* endotype, and in section 4.3 we find that *Streptococcus* relative abundance in *non-ARI* samples from the first seven *weeks-of-life* to be predictive of it. The sample sizes in this data set are very small, but the Mann-Whitney U-test still found the predictive power significant (p-value $< .05$).

Finally, a detailed extraction of how the relative abundances were obtained is provided in appendix F.

4.2 Inference of predictor and genera interaction from dual qq-plots

4.2.1 Results from ARI-*data in the first 90* days-of-life

We found that restriction to seven-week *ARI* samples had insufficient statistical power to generate useful dual qq-plots, so we used *ARI*-samples from the first 90 *days-of-life*. We also sought further support for important *genera* by testing their ability to predict their associated outcome. Such a test is prone to false negatives because of its linearity assumptions, but it is gratifying when such a signal is found. Microbiome samples were taken with every respiratory infection, so some individuals contributed multiple *ARI* samples from the first 90 *days-of-life*. We dealt with this issue by using the h-block cross-validation [172]. This was not a consideration for *non-ARI* samples from the first seven *weeks-of-life* because no individual gave more than one such sample.

Haemophilus in *ARI*-samples is associated with increased risk of *wheeze via* a *non-atopic* mechanism

Figure 4.2 indicates that the ARI samples of those who go on to exhibit *wheeze* in the fifth *year-of-life* are associated with higher relative abundances of *Haemophilus*. The corresponding result is less clear for the specific fifth-year endotype *atopic-wheeze* (figure

E.1) but is stronger for *nonatopic-wheeze* (figure 4.3). This associates *Haemophilus* with a *non-atopic* mechanism of *wheeze*, although we cannot comment on whether it is actually causal.

Moraxella in *ARI*-samples is associated with a decreased risk of *atopic-wheeze* in the fifth *year-of-life*

The dual qq-plots in figure 4.4 show an apparently protective effect of *Moraxella* against *atopic-wheeze*, where those who developed fifth-year *atopic-wheeze* were associated with lower relative abundances of *Moraxella*. This held to a lesser extent for *wheeze*, as shown in figure E.2 but not for those who went on to develop fifth-year *nonatopic-wheeze*, for whom the relative abundance of *Moraxella* is higher than for those who did not (see figure 4.5). We cannot determine if *Moraxella* relative abundance actually led to *nonatopic-wheeze*, which seems unlikely given its association with reduced *wheeze* risk, but it follows that it was associated with lowered risk for *atopic-wheeze*.

Moraxella from these samples was a statistically significant predictor of *nonatopic-wheeze* in the fifth *year-of-life* using logistic regression, with an AUC of .60 and a Mann-Whitney p-value of .01.

Staphylococcus in *ARI*-samples is associated with an increased risk of *atopic-wheeze* in the fifth *year-of-life*

The reverse is true for *Staphylococcus*, which represents a decreased risk relative to all other *genera* for *nonatopic-wheeze* (figure 4.6), but an increased risk for *atopic-wheeze* (figure 4.7). On the other hand, figure 4.1 shows that while a relatively high relative abundance of *Staphylococcus* in *ARI* samples taken during the first *year-of-life* (a functional cutoff would be 0.1) appears to protect against *nonatopic-wheeze* in the fifth *year-of-life*, the counter-examples all came from a specific individual who later submitted four other *ARI* samples in which the relative abundances of *Staphylococcus* were well below this cutoff. This is not consistent with elevated *Staphylococcus* actively contributing to *atopic-wheeze* in the fifth *year-of-life* and suggests an indicative relationship, although it is also possible that some other factor has arrested the pathogenic

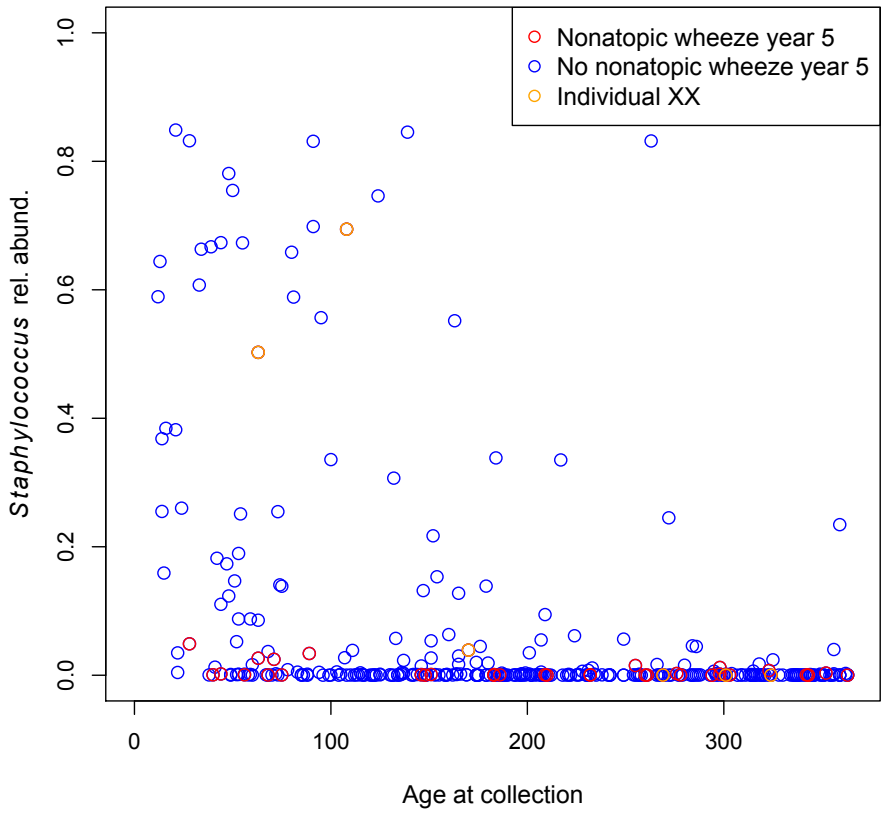


Figure 4.1: **Relative abundance of *Staphylococcus* in biome samples taken from participants during a respiratory infection in the first *year-of-life*. Samples with a high relative abundance of *Staphylococcus* did not come from those who later exhibited *nonatopic-wheeze* in the fifth *year-of-life*, with two exceptions whose common donor, whose samples are labelled XX in the plot, later yielded *ARI* samples with low relative abundances of *Staphylococcus*, suggesting the relationship is probably indicative:** Since *Staphylococcus* in early (90 days) *ARI* samples was a risk factor for *atopic-wheeze* in the fifth *year-of-life*, we conclude that it was not so much protective as associated with the *atopic* asthma endotype instead.

process.

We are reminded of Barnes' suggestion [28] that non-allergic asthma might be driven by so-called *superantigens* produced by *Staphylococcus* and similar microbes, but our results indicate that *Staphylococcus*, at least when accompanied by respiratory infection, is a risk for *atopic-wheeze* instead.

NP *genus*	*wheeze*	*atopic-wheeze*	*nonatopic-wheeze*
Streptococcus	**.043**	**.016**	.953
Streptococcus and *Haemophilus*	**.044**	**.015**	.958

Table 4.1: **Mann-Whitney p-values of *wheeze* dependence on the relative abundance of *Streptococcus* in *non-ARI* samples taken during the first seven *weeks-of-life* (logistic regression).** We considered two classes of model, one in which only *Streptococcus* was considered and one in which *Haemophilus* was considered also.

Both fifth-year *nonatopic-wheeze* and *atopic-wheeze* were predicted with the logistic regression classifiers, scoring AUCs of .67 and .66, respectively. (The AUC was .65 for both endotypes using the naïve Bayes classifier.) The AUC was significant only for *atopic-wheeze* with a Mann-Whitney p-value $\approx .03$.

4.2.2 Fifth-year wheeze *from 7-week non-*ARI *data*

The dual qq-plots in figure 4.8 confirm that samples from participants who go on to exhibit *wheeze* in the fifth *year-of-life* are associated with higher relative abundances of *Streptococcus*. This is at the expense of every other *genus* except *Haemophilus*. Although these signals look significant to the eye, we quantified their significance where we could. We used logistic regression to model the dependence of *wheeze* in the fifth *year-of-life* on the relative abundance of *Streptococcus* and other *genera*. It was not appropriate to include the relative abundances of all *genera*, since they are constrained to sum to one, so we found two models, one with *Streptococcus* only and another that also included *Haemophilus*. We considered *Haemophilus* because it is the only *genus* for which figure 4.8 shows no antagonism with *Streptococcus* . The results are shown in table 4.1. Unlike the dual qq-plots there is a stronger signal for *atopic-wheeze* than for *wheeze* in general, associating *Streptococcus* in these samples with an *atopic* mechanism. While the calculated significance is only marginal, it is reasonable to consider this an underestimate of significance since logistic regression assumes a log-linear relationship, which is a poor assumption in this case. Values for *Haemophilus* are not shown as it was never found to have a significant coefficient.

4.3 Prediction of allergen-specific *atopic-wheeze* from microbiome relative abundance in infant microbiome data

To see if the microbiome relative abundances were endotype-specific we studied their predictive power for *allergen-specific atopic-wheeze* in the fifth *year-of-life*. We found some suggestive results for *Streptococcus*, though we lacked the statistical power to make any definitive claims.

The relative abundance of *Streptococcus* in *non-ARI* samples from the first seven-weeks of life scored an AUC of .60 from both classifiers predicting fifth-year *atopic-wheeze*, but scored AUCs in the mid-to-high 80% range, predicting almost all of the allergen-specific *atopic-wheeze* endotypes identified in chapter 3. The exception was *rye*-specific *atopic-wheeze* whose predictivity was not statistically significant. The corresponding AUCs and p-values are presented in table 4.2. (The AUCs given in this chapter were calculated using logistic regression, however naïve Bayes gave the exact same AUCs for all cases in this chapter except for prediction of *multi-allergenic atopic-wheeze* by *Streptococcus* relative abundance in *non-ARI* samples from the first seven *weeks-of-life*. In this the naïve Bayes classifier scored less than logistic regression by .025)

While the *multi-allergen* endotype was also well-predicted (AUC=.76), the poor prediction of *rye*-specific *atopic-wheeze* prompted us to consider a reduced form where *rye* was not included. While this did score an improved AUC of .83, the result is hardly surprising and the exclusivity index from equation (3.5) was $.58 \pm .02$, which does not support an exclusive subtype. In case this was due to the limited statistical power of the microbiome samples we repeated the calculation using (log of) *house-dust-mite IgE* from the second *year-of-life* and again found no evidence, scoring an index value of $.76 \pm .19$. (For the interested reader, the corresponding AUC was .96 instead of .94, with an identical p-value of 7.8×10^{-14} from a smaller number (18) of cases.)

The relevant p-values (table 4.2) are less than .05 but not overwhelmingly so. Even if the significance cutoff did not need to be adjusted for six *genera* and six relevant

response atopy	AUC	p-value	nb. cases
cat	.84	**.024**	5
peanut	.87	**.035**	3
couch	.89	**.007**	4
rye	.66	.15	6
cat/peanut/couch/rye	.76	**.022**	8
cat/peanut/couch	.83	**.006**	7
house-dust-mite	.62	.19	11

Table 4.2: **AUCs of *atopic-wheeze* endotypes from the relative abundance of *Streptococcus* in 7-week non-infectious microbiome samples:** The relative predictability of the specific *atopies* is consistent with *Streptococcus* being exclusively predictive of the *multi-allergen* endotype identified in section 3.4, except for *rye*. The sample sizes were very small, as indicated, and combined with the overtesting issues discussed in the text, we present this as a strong hint for *Streptococcus* being exclusively predictive of this endotype, but lack the statistical power to claim this as a result. Again, p-value is found from the Mann-Whitney test.

allergens (*house-dust-mite*, *peanut*, *mould*, *couch*, *rye* and *cat*), the corresponding sample sizes would not allow us to confidently identify an endotype on this basis. We simply wish to note that the AUCs are strongly suggestive of *Streptococcus* being relevant to at least some cases of the *multi-allergen* endotype inferred from the exclusivity index (equation (3.5)) in section 3.4. This must remain a suggestion for now as we lack the statistical power to take it further.

4.4 Outcomes

The relevance of the NP microbiome to *wheeze* and early infection had already been demonstrated by [7]. In this chapter we have used dual qq-plots to demonstrate new associations with later *wheeze* endotypes, and to provide additional detail on the association of *Streptococcus* in *non-ARI* biome samples and *wheeze* in the fifth *year-of-life*. Questions concerning causality remain. None of the associated *genera* seem to be necessary or sufficient, so it is not clear that they are actually causal. Tools such as Pearl's causal calculus [101] can sometimes shed light on such questions but that is beyond the scope of this work. Nonetheless, we may still conclude that at least some causal factors are manifest at this very early time of life. A summary of this chapter's outcomes is as follows:

1. The relative abundance of *Haemophilus* in *ARI* microbiome samples during the first 90 *days-of-life* was associated with *wheeze* in the fifth *year-of-life via* a *non-atopic* mechanism. This was not detectable with a logistic regression model.

2. The relative abundance of *Moraxella* in *ARI* microbiome samples during the first 90 *days-of-life* was associated with a reduced risk of *wheeze* and especially *atopic-wheeze* in the fifth *year-of-life*, but a greater risk of *nonatopic-wheeze*. This last association was also detected by a logistic regression classifier.

3. The relative abundance of *Staphylococcus* in *ARI* microbiome samples during the first 90 *days-of-life* was the converse of *Moraxella*, with greater risk of *wheeze* and *atopic-wheeze* and a reduced risk of *nonatopic-wheeze*. This would indicate a fifth-year *wheeze* risk *via* an *atopic* mechanism.

4. The above associations of *Staphylococcus* are most likely indicative, since the individual who developed *nonatopic-wheeze* in the fifth *year-of-life* after donating (two) *ARI* samples with high relative abundances of *Staphylococcus*, submitted later ones with low relative abundances of *Staphylococcus*.

5. The relative abundance of *Streptococcus* in *non-ARI* microbiome samples during the first seven *weeks-of-life* was associated with greater risk of *wheeze* in the fifth *year-of-life* according to the dual qq-plots. This effect was detectable by logistic regression, which also detected an increased risk of fifth-year *atopy* with greater significance. However the corresponding effect on *atopic-wheeze* is not visible in the dual qq-plots.

6. The relative abundance of *Streptococcus* in *non-ARI* microbiome samples during the first seven *weeks-of-life* was scored relatively high AUCs for almost the same allergen-specific *atopic-wheeze* that (log of) *house-dust-mite IgE* from the second *year-of-life* was predictive of. The exception was *rye*-specific *atopic-wheeze*, but the exclusivity index did not find this to be indicative of a more precise endotype. It is suggestive of *Streptococcus* in these samples being important for *multi-allergen*

atopic-wheeze, but we concede that there is insufficient statistical power to make a definitive claim.

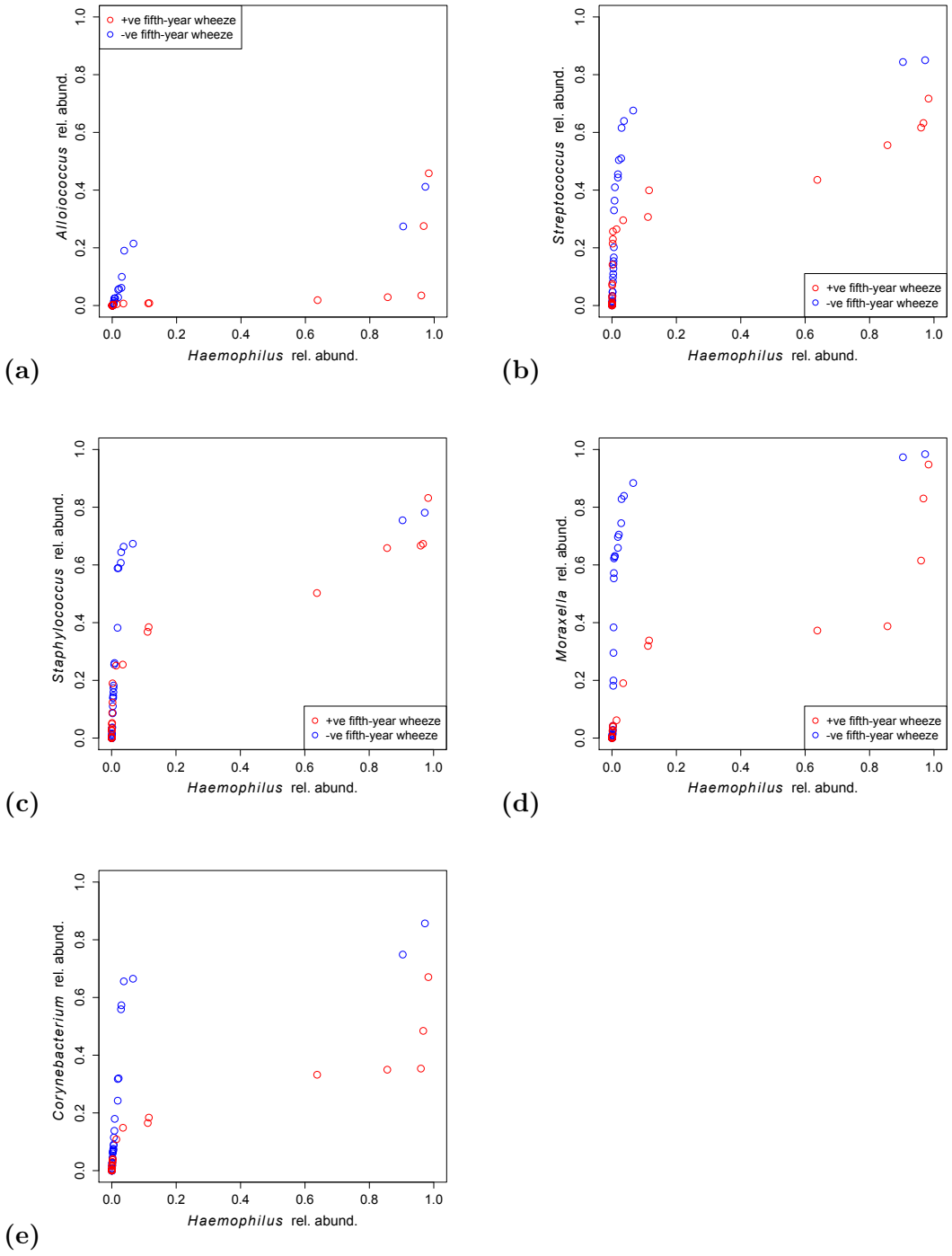


Figure 4.2: **QQ-plots demonstrating a link between an elevated relative abundance of *Haemophilus* in biomes from the first 90 *days-of-life* when the participant was suffering a respiratory infection, and *wheeze* in the fifth *year-of-life*:**

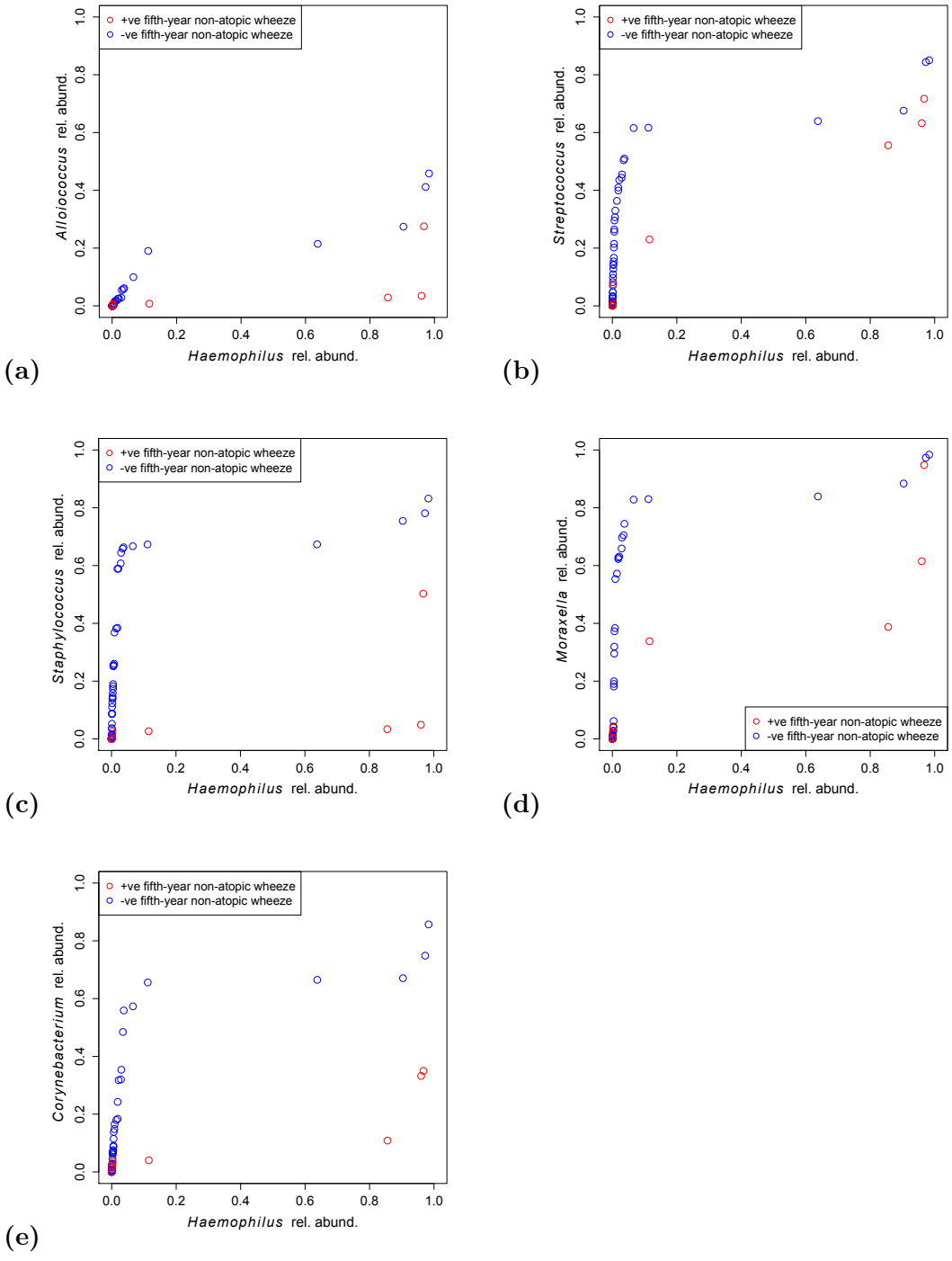


Figure 4.3: **QQ-plots demonstrating the link between an elevated relative abundance of *Haemophilus* in *ARI* biomes from the first 90 *days-of-life*, and *nonatopic-wheeze* in the fifth *year-of-life*:**

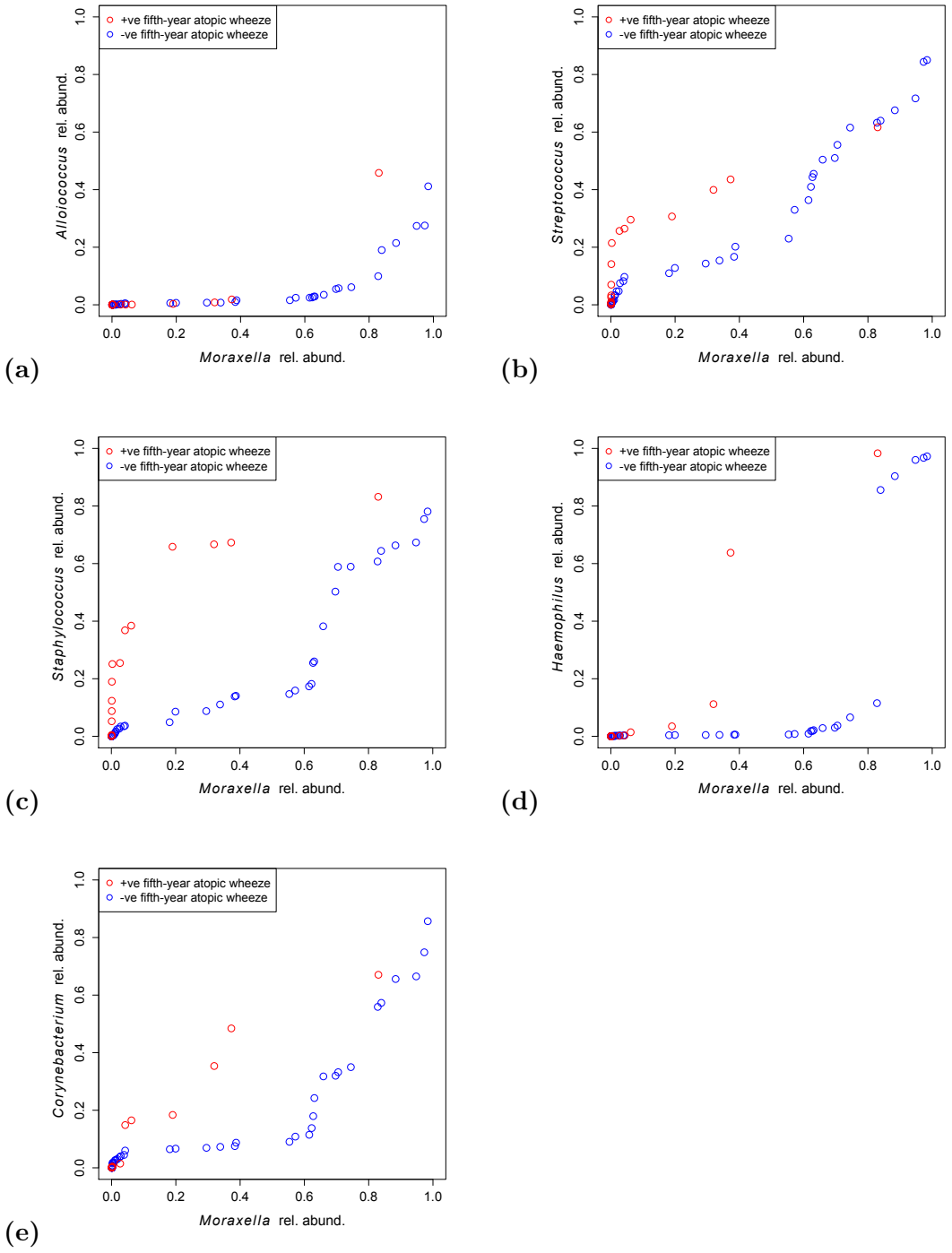


Figure 4.4: **QQ-plots demonstrating a link between an elevated relative abundance of *Moraxella* in biomes from the first 90 *days-of-life* when the participant was suffering a respiratory infection, and lack of *atopic-wheeze* in the fifth *year-of-life*:**

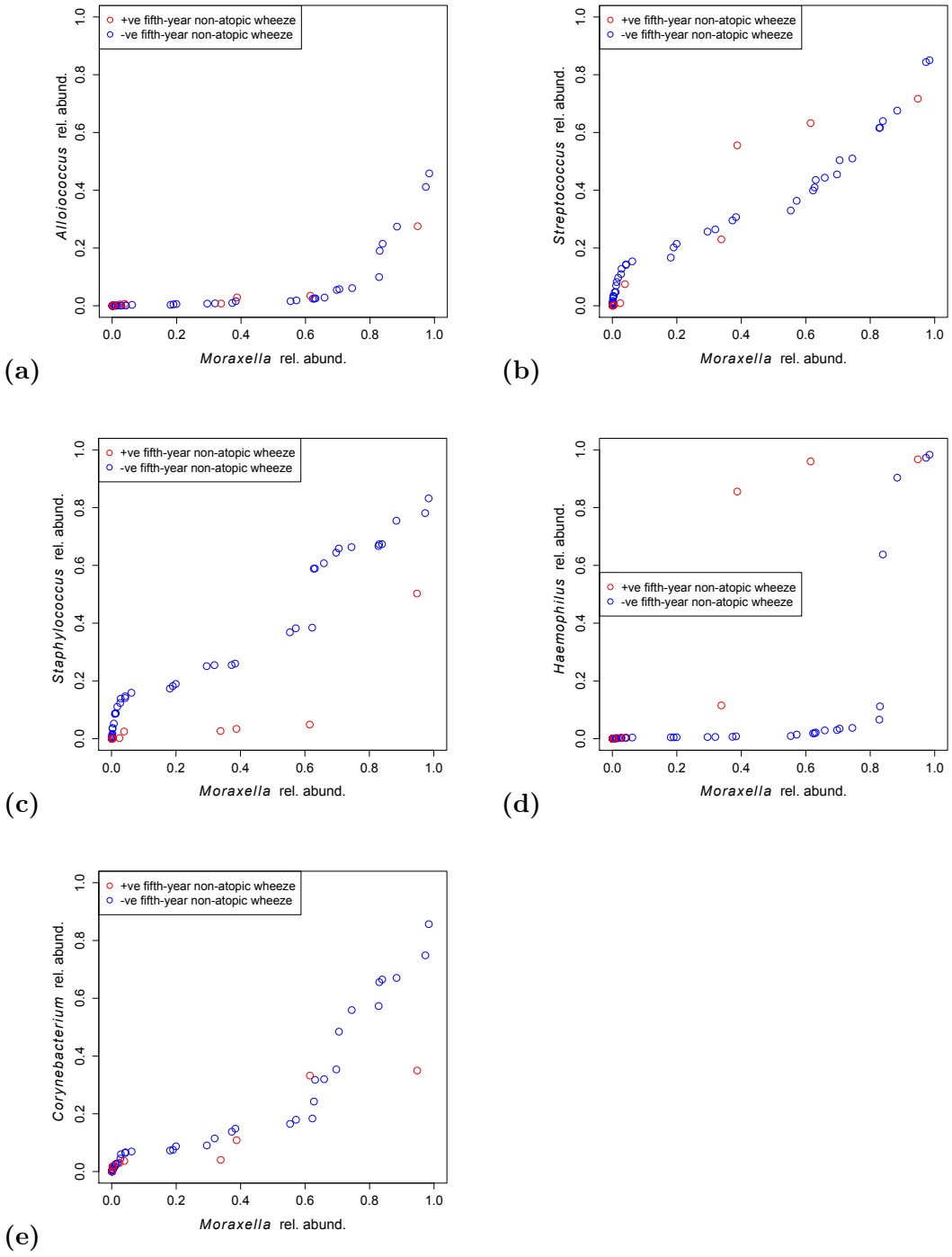


Figure 4.5: **QQ-plots of the relative abundance of *Moraxella* compared to other *genera*, where we distinguish between samples that are postive or negative for *nonatopic-wheeze* in the fifth *year-of-life*:** The protective effect observed against *atopic-wheeze* in figure 4.4 is absent.

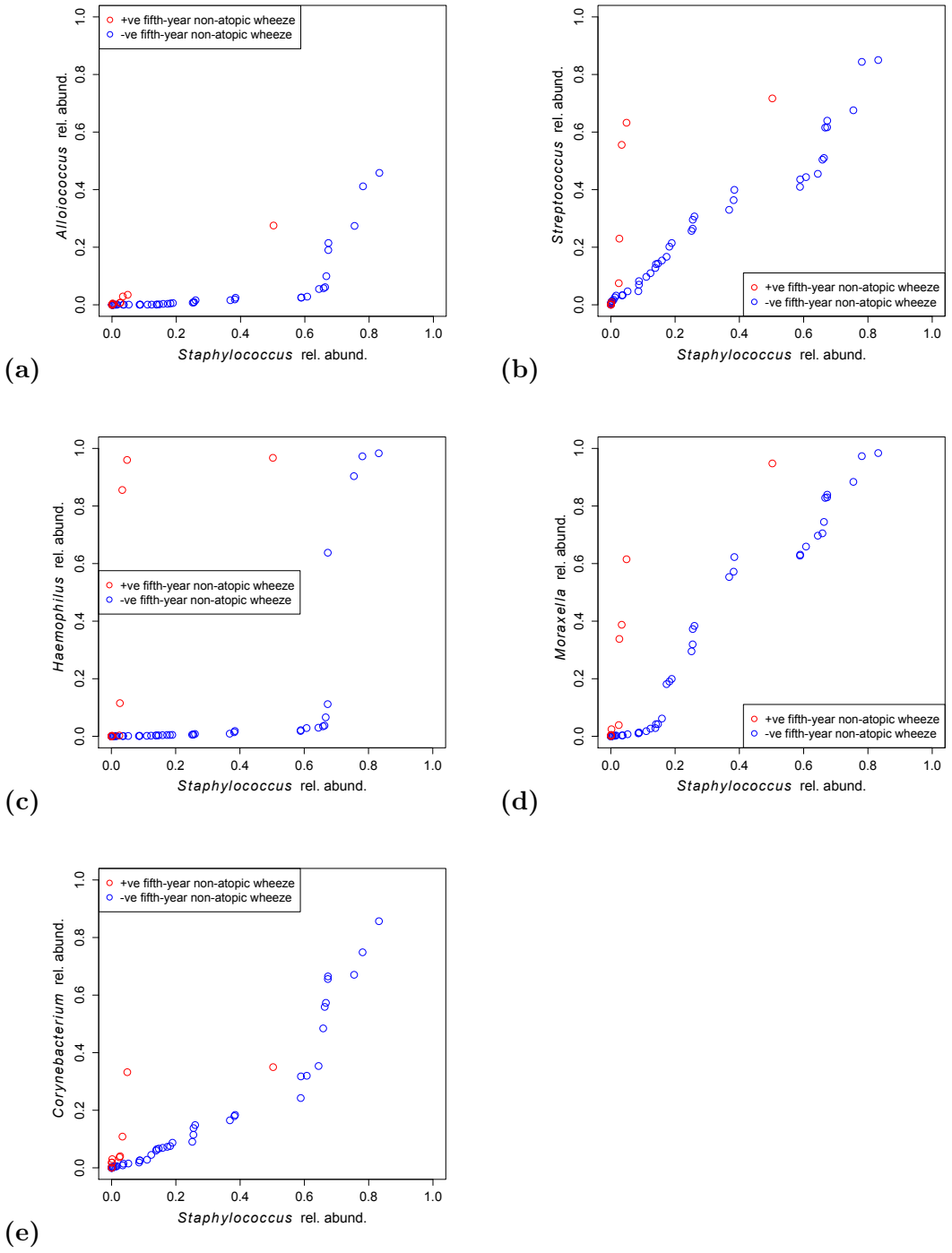
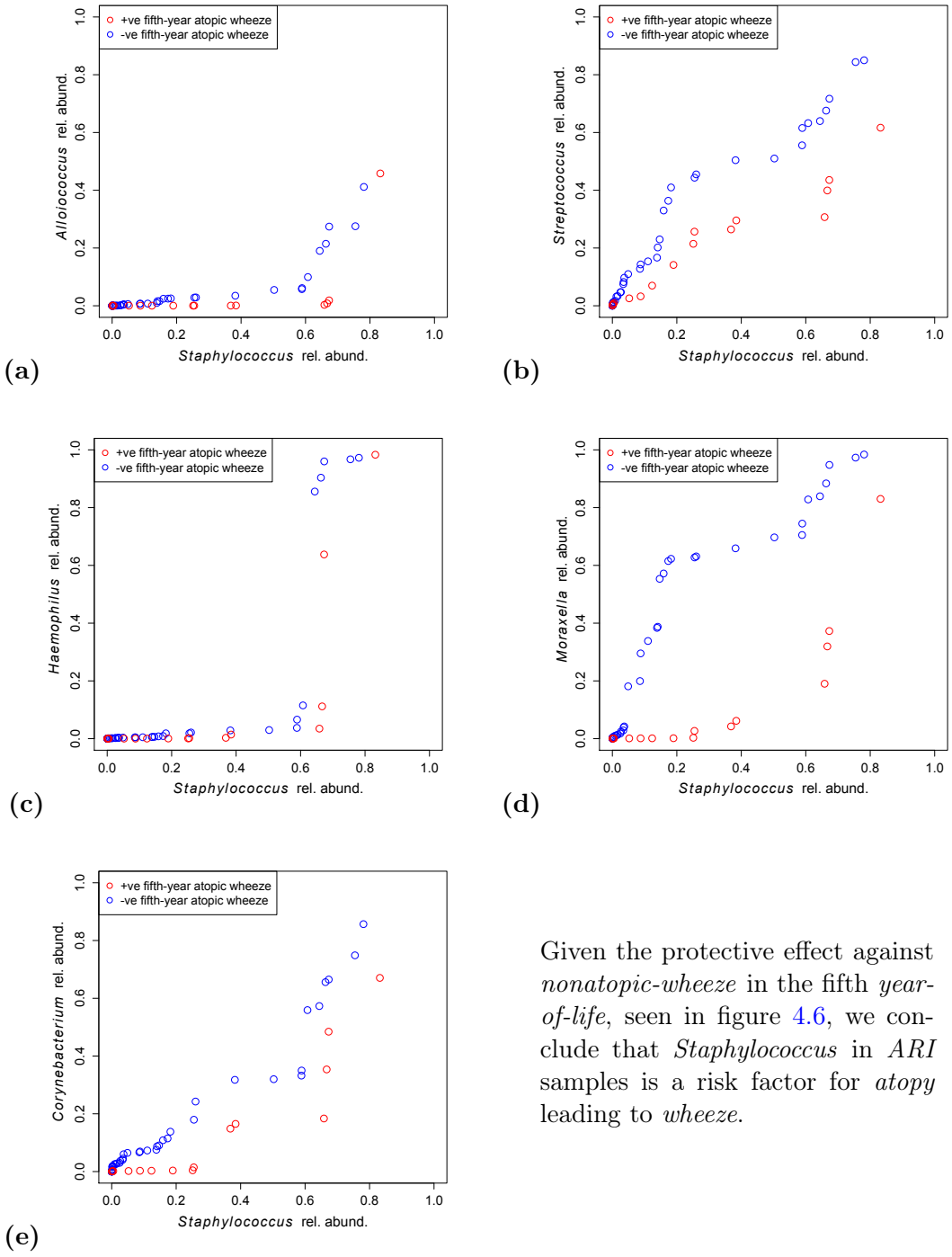


Figure 4.6: **QQ-plots demonstrating a link between an elevated relative abundance of *Staphylococcus* in biomes from samples taken during the first 90 *days-of-life* when the participant was suffering a respiratory infection, and lack of *nonatopic-wheeze* in the fifth *year-of-life*:**



Given the protective effect against *nonatopic-wheeze* in the fifth *year-of-life*, seen in figure 4.6, we conclude that *Staphylococcus* in *ARI* samples is a risk factor for *atopy* leading to *wheeze*.

Figure 4.7: **QQ-plots demonstrating a link between an elevated relative abundance of *Staphylococcus* in biomes from the first 90 *days-of-life* when the participant was suffering a respiratory infection, and *atopic-wheeze* in the fifth *year-of-life*:**

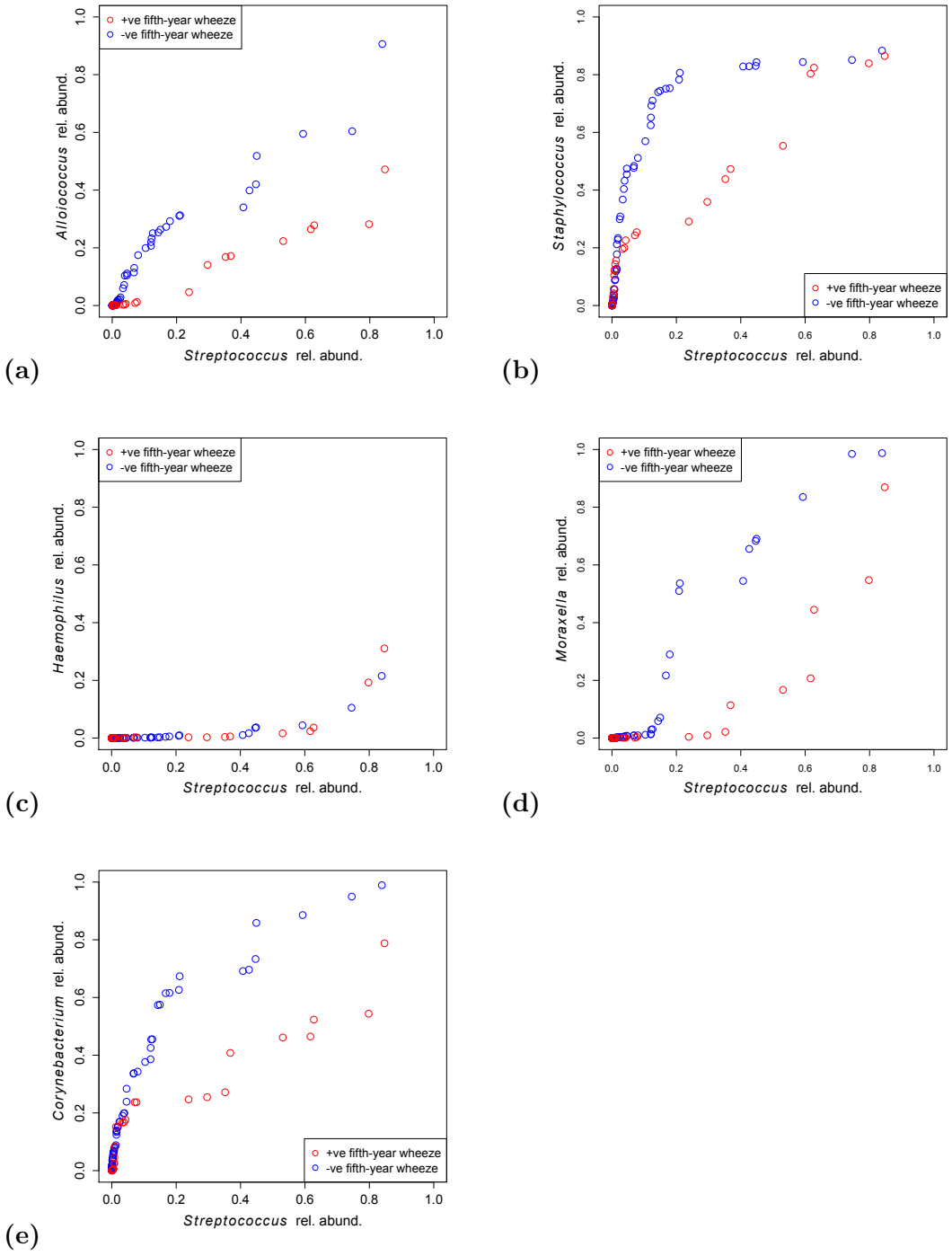


Figure 4.8: **QQ-plots demonstrating a link between an elevated relative abundance of *Streptococcus* during the first seven *weeks-of-life* in biome samples taken when the participant was free of respiratory infection, and *wheeze* in the fifth *year-of-life*:**

Chapter 5

Inferred dynamic Bayesian networks

5.1 Introduction

We have seen that the causes and aetiology of asthma are complicated. In chapter 3 we saw that important factors sufficient to define a biological endotype of *wheeze* may well be restricted in their relevance to such a small minority of asthma cases that their predictive ability for the general condition is swamped by noise. As such they cannot be used by simple methods being applied to predictions of *wheeze*. In chapter 4 we saw that at least some causal factors are present even in the first seven *weeks-of-life*. The many variables of asthma pathogenesis require non-trivial networks to understand the interactions amongst them.

We chose our network variables to address questions relevant to the current state of the field, often beginning with general variables, say *LRI*, before moving on to more specific variables, such as *mild-* and *severe- LRI*. Much of our effort was spent studying networks of infection, *wheeze* and *atopy*. We also found interesting networks among *IgE*s and interleukins.

5.2 Technical considerations concerning our inference of dynamic Bayesian networks

Rather than reinvent the wheel, we used the *ARTIVA* package, originally created to infer genetic interaction networks [8], to generate our DBNs. We have already described how *ARTIVA* works in our review of Bayesian networks (subsection 2.2.1). In this section we discuss our handling of the technical issues. Since Lebre *et al.* discuss the technicalities of *ARTIVA* quite thoroughly, we largely accept their default parameters, including the number of iterations, for which their default was 50,000, more than enough for the

RJ-MCMC algorithm to "burn-in" [8]. We allowed the maximum possible number of changepoints, which was one less than the number of time-points. We set *edgesThreshold* to .4 and took the default values for the other input variables. We used R version 3.1.1. Unfortunately *ARTIVA* is no longer maintained by its authors so care must be taken in upgrading to a more recent version of R.

5.2.1 Independence approximation and choice of variables

The *ARTIVA* algorithm assumes that all parent nodes are independent. Since many of the variables are confounding simply by virtue of their relevance to asthma, it was impossible to ensure that this was even approximately true in many of our networks. Although Bayesian methods have often been found to tolerate the violation of this assumption remarkably well [124–126], this was nonetheless a strong incentive to avoid overly large networks and instead choose networks of specific variables to address specific questions.

For this reason we also chose variables whose definitions did not overlap. For example, we have avoided including *severe-LRI* with either *wheezy-LRI* or *febrile-LRI* because the latter two are special cases of the first, although we sometimes commit minor violations of this rule since *wheezy-LRI* and *febrile-LRI* share an overlap. This was generally not a problem but some networks were improved by distinguishing the purely *febrile* or *wheezy* infections from those that were both (*e.g.* see figure 5.22).

Conversely, since we wish to ascertain which specific variables are relevant, it was also appropriate to cover all subtypes of a given variable, and not just the ones for which we expected an effect. For example, in figure 5.15 when we expected *wheezy-LRI* to be the only *severe-LRI*s that interacted we still included *febrile-LRI* to be sure that the set was covered. Another illustrative example is figure 5.18, where we included *severe-LRI* with *nonairborne-atopy* when considering the importance of accompanying *severe-LRI* with *atopy* but wanted to focus on *airborne-atopy*.

5.2.2 Augmented discrete variables in Gaussian process priors

As explained in the review of Bayesian methods 2.2, *ARTIVA*'s Gaussian prior is ill-suited for discrete data. Since some of CAS' most important data, such as infection count data and multiple binary variables are discrete, we implemented a much-cited solution first proposed by Albert and Chibb [9], in which discrete data are augmented by latent variables which are sampled from truncated Gaussian distributions just as the linear coefficients are.

Since this procedure need only be applied to data at the child nodes, and *ARTIVA* works by inferring parents for each child, it was straightforward to do this consistently. We also used the fact that the discrete data was either count or binary data to simplify Albert and Chib's procedure. The complete procedure for arbitrary discrete data includes a step in which the cutoffs between discrete categories are also sampled [9]. This is necessary for categorical data in general because the integer labelling of categories is entirely arbitrary. This is not so for the special case of count data however, because the categories "zero", "one", "two" *etc.* meaningfully correspond to the integer values 0, 1, 2 *etc.*, allowing us to neglect this step.

Our implementation is contained in a module which is called from within *ARTIVA*'s module "main.R". The code is available by request from the author at
m.walker@aip.org.au .

5.2.3 How we rendered our networks

The layout of our DBNs

The graphs representing each network have one node for each variable at each time-point. These are laid out in a grid. The rows correspond to the variables listed along the y-axis, while the columns correspond to the time-periods, usually the first to the fifth *years-of-life*, although some networks show data at six months or at birth (from cord blood), labelled by "6m" and "birth" respectively.

The effect strength (width) and sign (dashed is negative) associated with edges

The presence of an edge indicates that the value of the child node varies (approximately) linearly with that of the parent node. The width is proportional to the corresponding linear coefficient, with solid lines indicating a positive coefficient and dashed lines indicating a negative coefficient. We describe this in the text by saying that a parent variable *led to* a child variable if there is solid edge connecting the corresponding nodes, and that it *led against* if the effect is a negative one (dashed line). If a variable led to itself from one year to the next we describe it as ongoing. We also capped the width of the lines at two in order to keep the graphs legible.

The posterior probability of an edge determines its shade

As discussed above, *ARTIVA* returns the posterior probability of each edge which we found to be valuable in identifying appropriate variable choices and subsets. While all edges are coloured black, their opacity fades with decreasing posterior, with the α-value equal to the edge's posterior.

When discussing the DBNs in the text, we often refer to variables acting on each other with a high, intermediate or low *posterior* (*probability*), and this corresponds to edges drawn in black or medium to light shades of grey. A legend showing the correspondence between shade and posterior probability is given in appendix G.

We shall argue in the following two subsections that edges with intermediate posterior indicate either a statistical limitation or that the relationship applies to only a subset of participants/events. We therefore used such *grey lines* to infer combinations of variables that are of more direct relevance, effectively allowing the data to guide us to better choices of variables.

Nodes corresponding to non-longitudinal data

Not all of the CAS variables were longitudinal. An important example of this was *transient wheeze*, a binary variable whose positive value indicated that *wheeze* was present in at least one of the first three *years-of-life* but absent in the fifth. Others

were one-off measurements, typically peri-natal. These included measures of breast milk fatty acid or protein content. Another example is the number of older siblings in the house at birth, found to be slightly predictive of fifth-year *nonatopic-wheeze* status in subsection 3.2.3.

Causality clearly requires that variables from the fifth *year-of-life* can only be child nodes. However, they might be particularly dependent on early-life data, say from the first two *years-of-life*. We were able to test for this by replicating these variables at multiple time points and instructing *ARTIVA* to only consider them as child nodes and never as parent ones. The child-only nodes of value to us were *transient wheeze*, whose definition includes fifth-year data, and *wheeze* in the fifth *year-of-life*. Child-only nodes do not appear in the first time-step interval of any graph, and are rendered in blue as a visual aid. The converse was true for peri-natal variables, which could only be parent nodes but might lead to later variables indirectly. We also had the capacity to replicate these variables at every time-step, but found no useful DBNs this way.

5.2.4 The effect of missing data

Missing data are to be expected in a longitudinal cohort study the size of CAS but *ARTIVA* does not tolerate missing data in its input, its creators having access to complete data sets [8]. This required us to either impute the missing data or restrict ourselves to complete records.

Mean imputation allowed us to retain more individuals, leaving us with greater statistical power for the variables that were not missing, but compromised sensitivity for those variables that did have a significant amount of missing data. Conversely, restriction to individuals for which none of the specified variables had no missing data allowed greater sensitivity for variables for which the data was missing for a significant number of individuals, but sometimes at the cost of weakening detection for other variables for which much less data was missing. The restriction approach also had the effect of restricting the number of variables in the networks, since each additional variable brought the risk of more missing data removing more individuals.

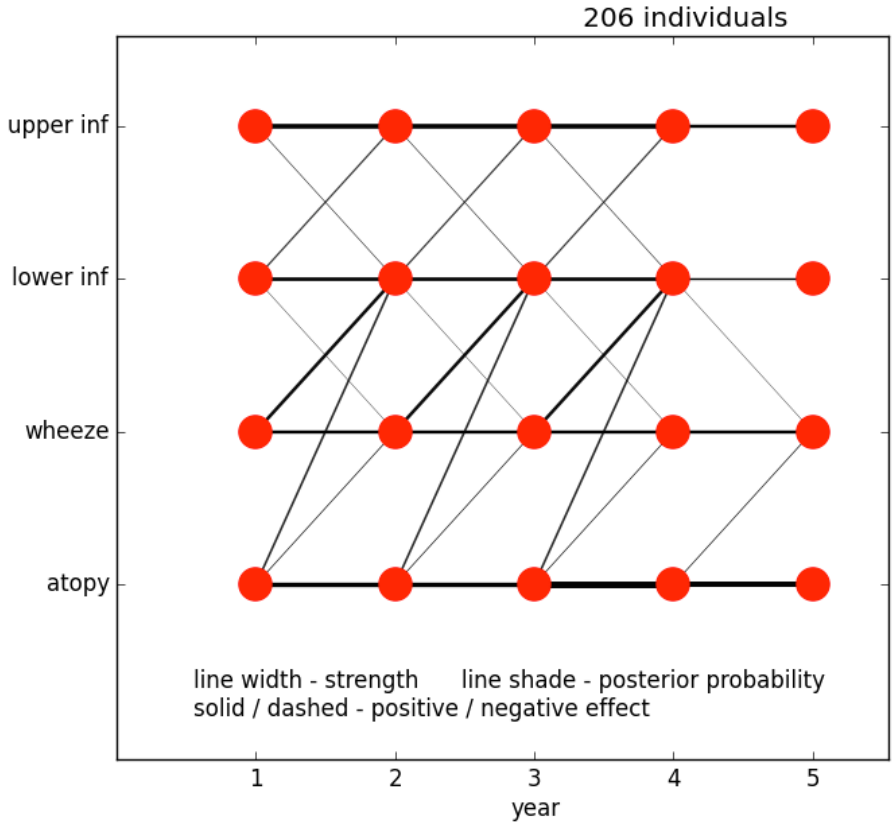


Figure 5.1: **DBN of *atopy*, *wheeze*, *LRI* and *URI* in which the missing data was mean-imputed:** There is a fading of edges leading to both *URI* and *LRI* in the fifth *year-of-life*. Otherwise, all four conditions appear to have been ongoing, with *wheeze* leading to *LRI* for the first three *years-of-life*. The same is true of *atopy*, though with only an intermediate posterior. This should be compared with figure 5.2.

The other problem with mean imputation is that it was incompatible with binary and count data, simply because the mean of such data is effectively never integral. Rounding off to the nearest integral was a potential source of error, giving us more reason to favour discarding individuals with missing data.

These issues are well illustrated by networks of general infection data. CAS had a lot of missing infection data, especially from the fourth and fifth *years-of-life*. The number of participants yielding *LRI* and *URI* data from the first to the fifth *years-of-life* was 194, 195, 189, 166 and 140, respectively. (This discussion is simplified by an exact correspondence between the participants lacking *LRI* data and those lacking *URI* data in any given *year-of-life*.)

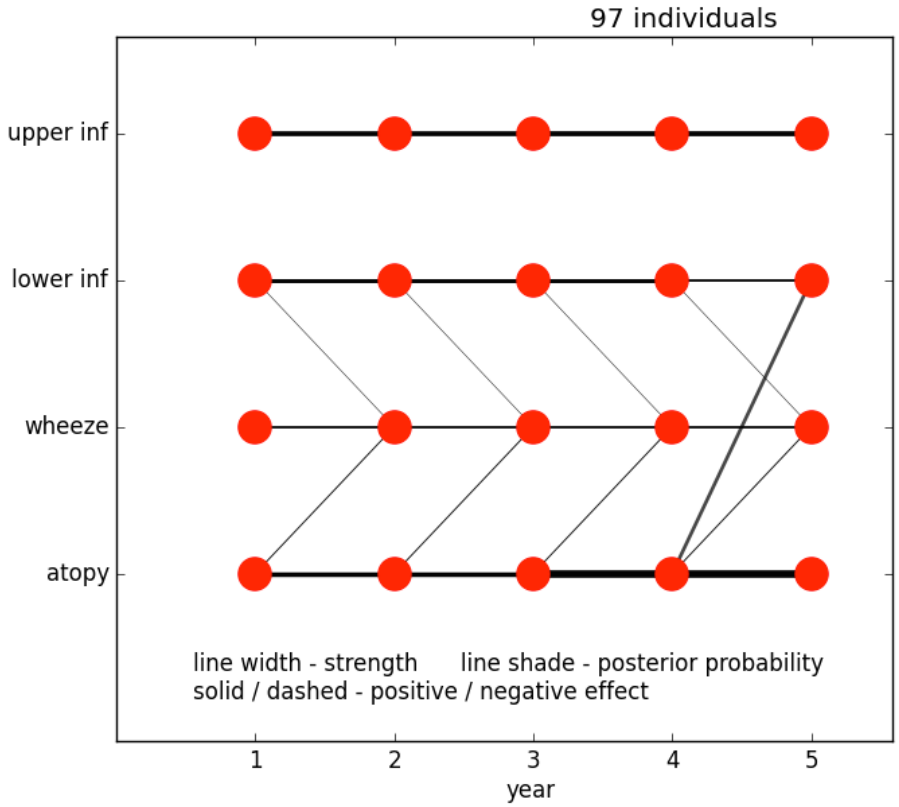


Figure 5.2: **DBN of *atopy*, *wheeze*, *LRI* and *URI* for which only complete records were used:** *LRI* does not show the same fading in the fifth *year-of-life* which was observed in figure 5.1. There is an additional edge to fifth-year *LRI* from *atopy*. On the other hand, earlier edges to *LRI* from *wheeze* and *atopy* are now missing.

We inferred two DBNs from *URI*, *LRI*, *atopy* and *wheeze*, one with mean-imputation and one by restriction to individuals whose records are complete in those variables. These networks are shown in figures 5.1 and 5.2, respectively. The number of data points available to infer an edge between the first pair (years one and two), second pair (years two and three), *etc.* finds sample sizes of 187, 183, 158, and 127 respectively. Since the overlap between these sets is imperfect, requiring all individuals to have complete data in these variables reduces the sample size to 97. Combined *wheeze* and *atopy* data, by contrast, suffer a much smaller attrition, with year-to-year sample sizes of 204, 202, 200, 186 and 169.

The DBN resulting from mean-imputation, shown in figure 5.1, indicates a fainter edge in the *LRI* row between the fourth and fifth *years-of-life* in comparison to earlier

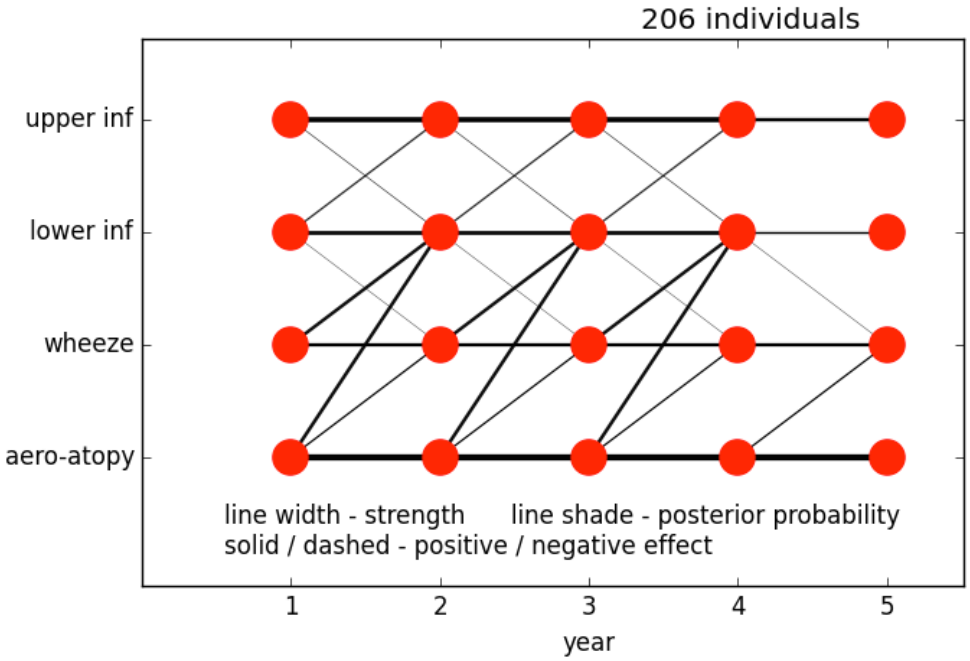


Figure 5.3: **DBN from replacing *atopy* with *airborne-atopy* in figure 5.1:** Edges from *airborne-atopy* are darker in this DBN than the corresponding edges from *atopy* in that of figure 5.1. This led us to the conclusion that *airborne-atopy* is a better choice of variable than *atopy*.

edges (legend in appendix G). The network inferred using only data from the 97 individuals with complete data for all four variables across all five years is shown in figure 5.2. The lower posteriors for the edges indicating the effects of *atopy* and *wheeze* are lower. In fact, the edges to *LRI* from *wheeze* and *atopy* during the first four *years-of-life* have vanished. Also, the edges connecting *URI*- and *LRI*- related nodes are unchanged except for that connecting the sequential *URI* nodes in the fourth and fifth *years-of-life*. That edge is now as dark and thick as the corresponding edges for earlier timepoints. This suggests that susceptibility to *URI* does not change going into the fifth *year-of-life*.

So mean-imputing missing data appears to 'wash-out' the corresponding signal, but protects connections with well-populated variables. The reader might question whether statistically meaningful results can truly be found from the smaller data set, but we have just seen that statistical uncertainty lowers the edges' posterior making false positives unlikely. It is worth remembering that the original application of *ARTIVA* generated networks from only ten samples [8].

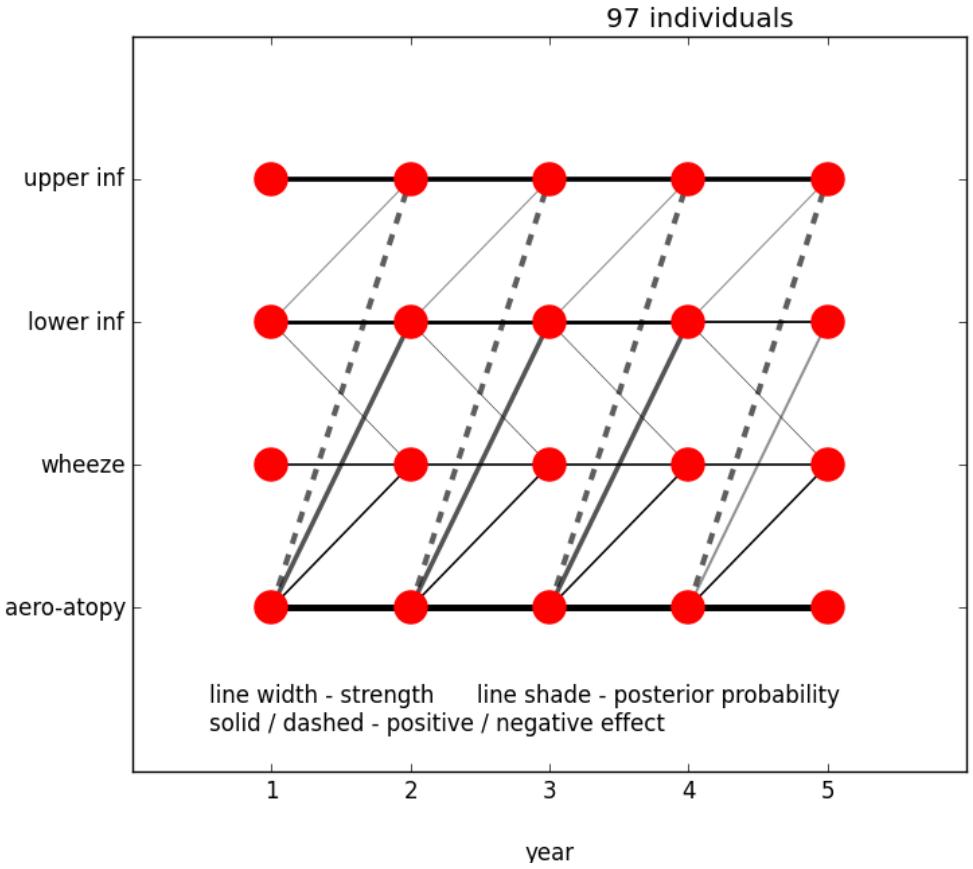


Figure 5.4: **DBN from replacing *atopy* with *airborne-atopy* in figure 5.2:** Similarly to figure 5.3, edges from *airborne-atopy* are darker in this DBN than the corresponding edges from *atopy* in that of figure 5.2. This again indicates that *airborne-atopy* is more important than *nonairborne-atopy*, and a better choice of variable than *atopy*.

Our approach was therefore to discard incomplete records for any given network, although the corresponding mean-imputed networks for infection variables sampled over five years are given in appendix H. The differences were typically small and similar in nature to the ones we have just discussed.

5.2.5 The meaning of intermediate posteriors

Some of the inferred networks have edges of intermediate posterior even when the corresponding data are not missing. Sometimes these edges are quite wide, indicating the inferred effect to be relatively large. Given sufficient statistical power, it is reasonable to attribute the reduced posterior to another 'wash-out' effect, due not to mean-imputed

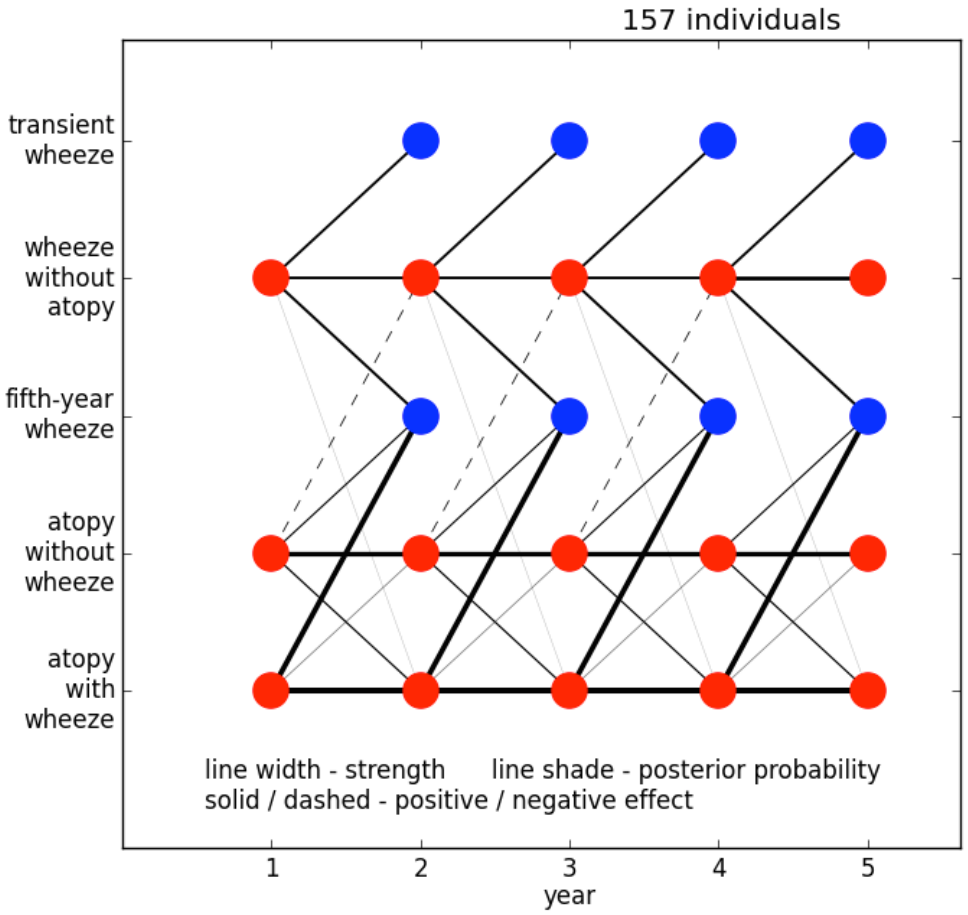


Figure 5.5: ***Atopy* with *wheeze* led to chronic (fifth-year) *wheeze*. Only *wheeze* without *atopy* could be transient. *Atopic* conditions were ongoing, *wheeze* without *atopy* less so:** *Atopy* without *wheeze* leads to *atopy* with *wheeze*. Unlike figure 5.6, there is also a faint edge going back.

missing data, but to the effect in question only being relevant to a subset of the events, such as particular types of a given infection, special cases of *atopy etc.* This is illustrated by comparing figures 5.3 and 5.4 to figures 5.1 and 5.2, respectively. We see edges leading to *LRI*, especially from *atopy* or *airborne-atopy*, have a higher posterior in the latter two figures than in the former two. In fact, edges from *atopy* to *LRI* vanish in figure 5.2. Since *airborne-atopy* is a subset of *atopy*, we take this to mean that the effect of *atopy* in these figures is due, or at least stronger for, the *airborne-atopy* subtype. The lower posterior results for *atopy* can then be attributed to a sort of wash-out effect where only a subtype (*airborne-atopy* in this case) contributes while its complement

predictor	*year-of-life*			
	1	2	3	4
wheeze with *airborne-atopy*	$\mathbf{9.26 \times 10^{-3}}$	$\mathbf{8.48 \times 10^{-4}}$	$\mathbf{2.87 \times 10^{-9}}$	$\mathbf{1.07 \times 10^{-7}}$
wheeze without *airborne-atopy*	.804	.343	.076	2.99×10^{-4}

Table 5.1: χ-squared test of *airborne-atopy* acting on fifth-year *wheeze*. Consistent with figure 5.6, the presence of *airborne-atopy* significantly increases the probability of *wheeze* persisting to the fifth *year-of-life*.

predictor	*year-of-life*			
	1	2	3	4
wheeze with *airborne-atopy*	.818	.806	.186	.069
wheeze without *airborne-atopy*	$\mathbf{2.95 \times 10^{-7}}$	$\mathbf{1.32 \times 10^{-3}}$	**.011**	**.921**

Table 5.2: χ-squared test of *airborne-atopy* leading to *transient wheeze*. Consistent with figure 5.6, the presence of *airborne-atopy* significantly decreases the probability of *wheeze* being transient.

(*nonairborne-atopy*) does not.

As a guard against hidden random effects, and to positively identify the relevant subtypes, such interpretations should be verified with a new network in which the subtypes of interest may be distinguished.

5.2.6 Independent testing of inferred edges with the χ-squared test

Bayesian approaches are adept at ignoring irrelevant data [113, 115], and are generally disinclined to find false relationships [114]. They also cope well with missing data [112], generally speaking. Taken together, these points indicate that while random effects such as biases in the missing data might conceivably weaken the apparent significance of a variable and/or one of its subsets, it would be highly unlikely to exaggerate the relevance of the complementary subtype. We have also already noted at the end of subsection 5.2.4 the generally low insensitivity of our infection-related networks to whether incomplete records are deleted or their missing data imputed. Nonetheless, we cannot assume flawless output from real-world data, so we also verified our important edges with the χ-squared test.

The χ-squared test and the *ARTIVA* algorithm, though reasonably consistent with each other, do not agree exactly. This is to be expected, since they are different models

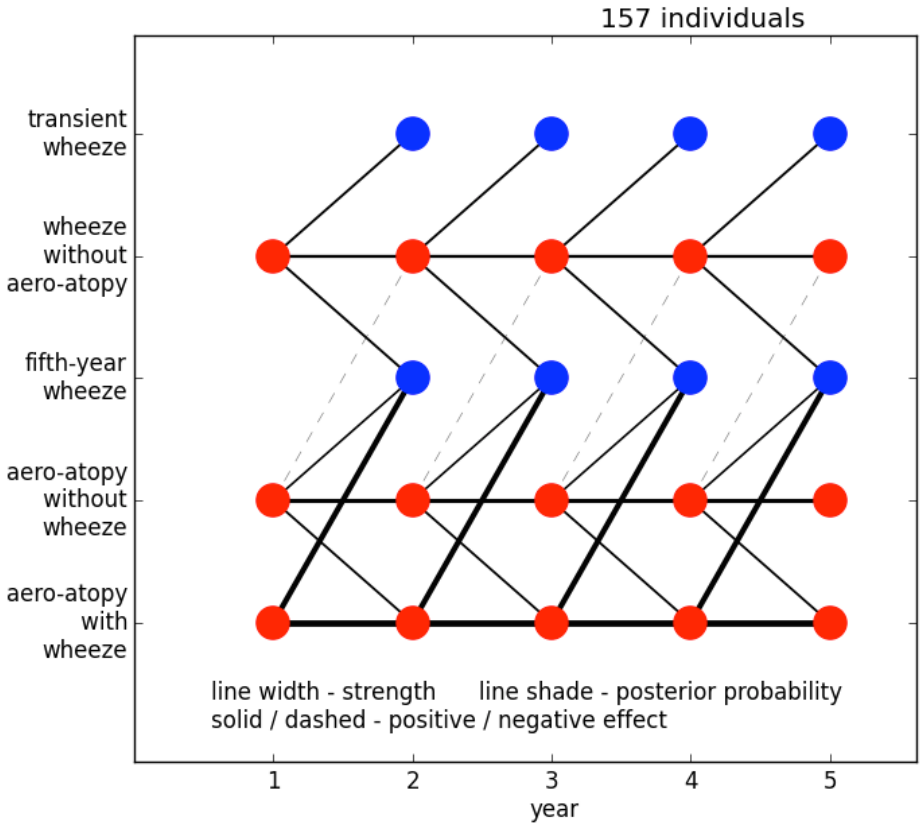


Figure 5.6: ***Airborne-atopy* with *wheeze* led to chronic (fifth-year) *wheeze*. Only *wheeze* without *airborne-atopy* could be transient. *Atopic* conditions were ongoing, *wheeze* without *airborne-atopy* less so:** *Airborne-atopy* without *wheeze* led to *airborne-atopy* with *wheeze*. Unlike figure 5.5, there is no edge going back the other way.

and include different information. The DBNs cannot describe relationships within a given *year-of-life*, although they may detect those which occur across sequential years, whereas the χ-squared test can. On the other hand, the χ-squared test cannot consider more than one pair of years at a time, while *ARTIVA* samples over an optimal range of years and can so avoid confounding effects associated with simultaneous changes over time between the two variables.

Tables 5.1 to 5.7 present the p-values from the χ-squared test. Where a link was predicted by *ARTIVA* it is presented in boldface, along with the p-value that was actually the smallest. We consistently found that the smallest p-value in each line of these tables corresponded to either the predicted edge or to values within the same year.

As such, the χ-squared test was always consistent with the inferred DBNs.

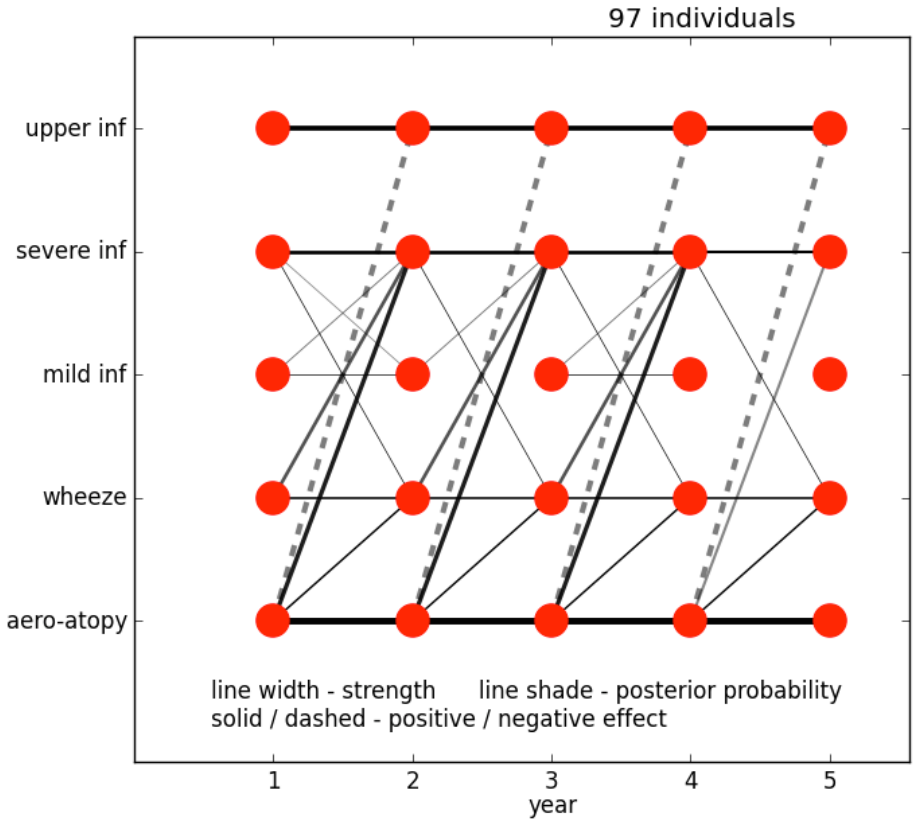


Figure 5.7: ***Airborne-atopy* led to *severe-LRI*. With intermediate posteriors, it also led against *URI* while *wheeze* led to *severe-LRI*:**

5.3 *Atopy*, infection, and persistence of *wheeze*

It is a long-established result that early childhood infections, especially febrile or wheezy infections, and viral infections, are associated with later childhood asthma. The same is true of childhood atopy, with 80-90% of asthmatics also suffering atopic (*IgE*-based) allergy [11, 19]. Figures 5.1 to 5.4 show *atopy* acting on the number of infections, rather than the other way around. There is an increase in *LRI* and, in the last of these figures (figure 5.4), a decrease in *URI*. The biologically plausible interpretation is that *aeroatopic-wheeze* makes it easier for infectious agents to get into the lungs, so that some infections which would have been recorded as *URI*s instead became *LRI*s.

It has long been known that *wheeze* commonly occurs in infants in the first two or three years of life only to resolve by about five years of age. Indeed, it is the

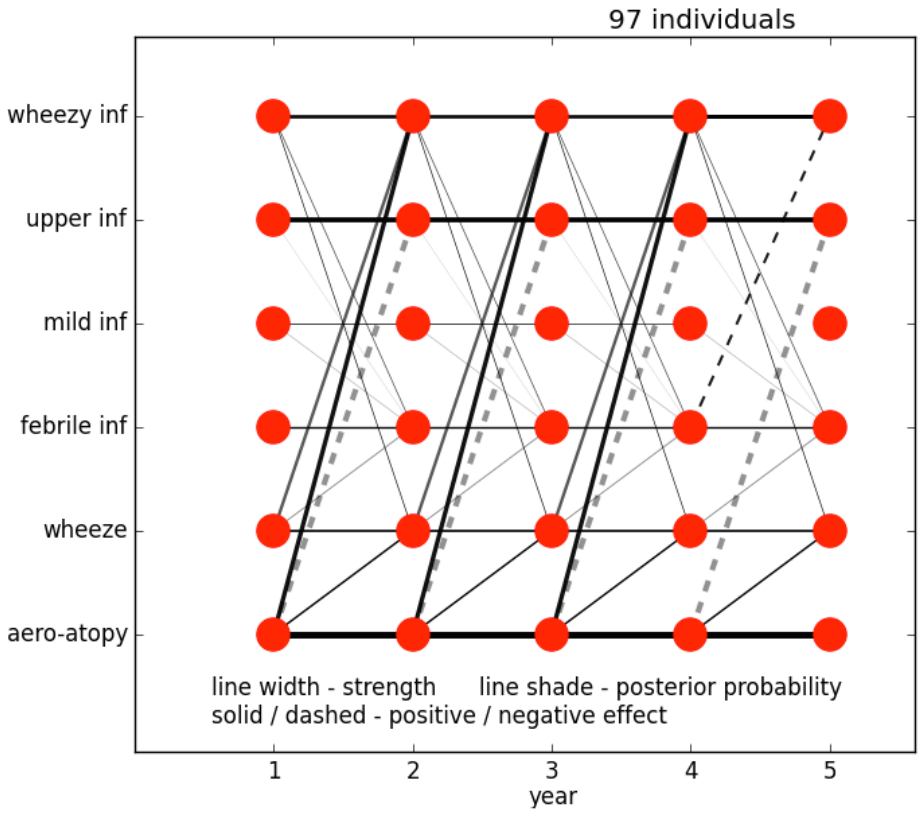


Figure 5.8: ***Airborne-atopy* led to *wheezy-LRI*. With low posterior it also led against *URI*:** There is a low-posterior link from *wheeze* to *wheezy-LRI*, which is not surprising.

persistence of *wheeze* into the fifth *year-of-life* that we have taken as our proxy for childhood asthma. Figure 5.5 indicates that the co-occurrence of *atopy* with *wheeze* increases the probability of the *wheeze* persisting into the fifth *year-of-life*. The effect is even stronger when *atopy* is restricted to *airborne-atopy* (figure 5.6), and vanishes when restricted to *atopy* to non-airborne allergens, or *nonairborne-atopy* (not shown). It would seem that the co-occurence of *airborne-atopy* is almost sufficient, but not necessary, to ensure that *wheeze* persists. Figures 5.5, 5.6 indicate that *airborne-atopy* could lead to *fifth-year-wheeze* on its own, but the effect is weaker, with lower probability (see legend in appendix G), and apparently *via atopic-wheeze* in a later year. Conversely, *wheeze-without-airborne-atopy* can persist to the fifth year-of-life but is more likely to be transient. The χ-squared test yields results consistent with this, as shown in tables 5.1 and 5.2.

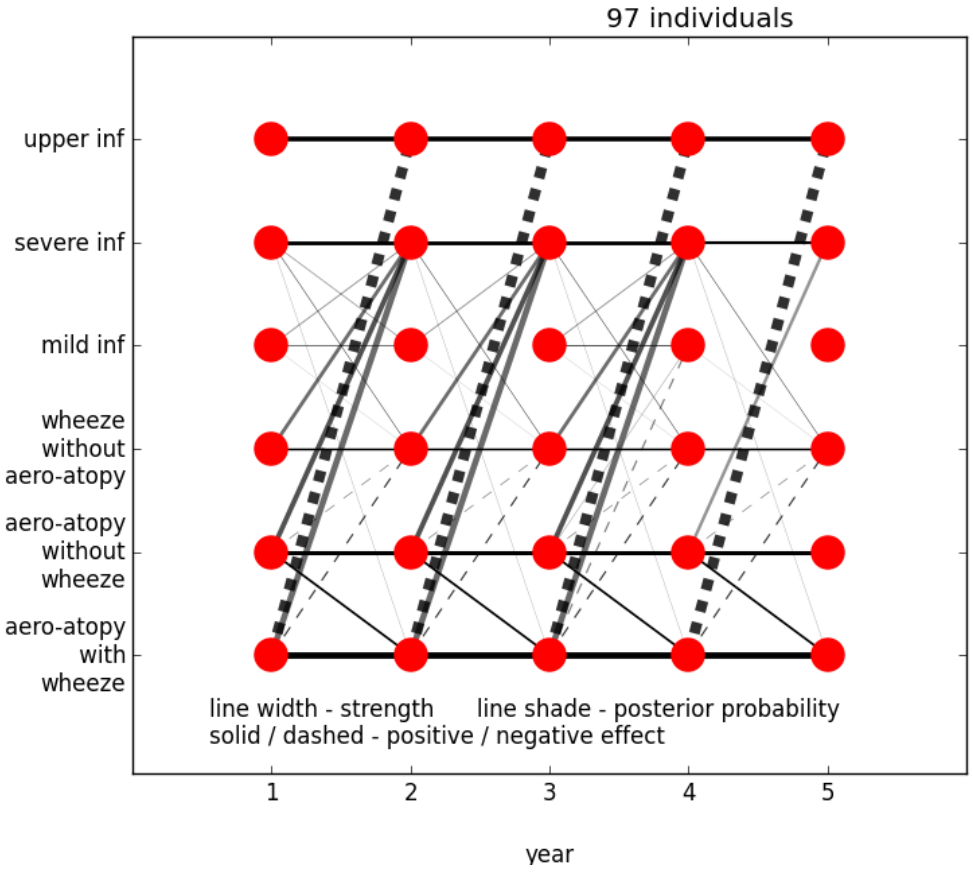


Figure 5.9: ***Airborne-atopy* with *wheeze* led against *URI*. *Airborne-atopy* and *wheeze* on their own led to *severe-LRI*:**

5.4 Did *LRI* lead to *atopy* and *wheeze*?

Since respiratory infections, especially lower respiratory infections, are associated with asthma risk, we put *URI*, *LRI*, *atopy* or *airborne-atopy*, and *wheeze* in the same network. As already mentioned, *airborne-atopy* seemed to lead *against URI*, as shown in figure 5.4. However, the posteriors of these effects are low-to-intermediate. As discussed in section 5.2.5, this might indicate that only a subset of infections are involved. Easy distinctions to make between infections include the copresence of *wheeze* and fever, and whether or not they are *viral*.

	aeroatopic-wheeze		
URI minus *severe-LRI*	previous year	same year	following year
URI, *wheezy-LRI* - yr 1	NA	**.045**	.183
URI, *wheezy-LRI* - yr 2	**.0103**	.047	.344
URI, *wheezy-LRI* - yr 3	**.635**	$\mathbf{6.69 \times 10^{-3}}$	.011
URI, *wheezy-LRI* - yr 4	$\mathbf{2.23 \times 10^{-6}}$	3.98×10^{-6}	1.75×10^{-4}
URI, *wheezy-LRI* - yr 5	$\mathbf{2.10 \times 10^{-4}}$	.022	NA

Table 5.3: χ-squared test p-values of the effect on the difference between the number of *URI*s and *wheezy-LRI*s from *atopic-wheeze*-status in the previous, same and following year. There is fair but incomplete agreement with the edges inferred in figure 5.11. Some effects are found here to occur on a timescale shorter than one year.

	atopic-wheeze		
viral-URI minus *severe-viral-LRI*	previous year	same year	following year
First *year-of-life*	NA	.172	.642
Second *year-of-life*	$\mathbf{1.98 \times 10^{-3}}$	5.83×10^{-3}	.387
Third *year-of-life*	**.414**	$\mathbf{1.86 \times 10^{-3}}$	.177

Table 5.4: χ-squared test p-values of the effect on the difference between the number of *viral-URI*s and *severe-viral-LRI*s from *atopic-wheeze*-status in the previous, same and following year. There is fair but incomplete agreement with the edges inferred in figure 5.14. Some effects are found here to occur on a timescale shorter than one year.

5.4.1 Atopy *and* wheeze *led to* LRI

To clarify the role of *LRI* we first decomposed it into *severe-* and *mild- LRI*. The resulting network (figure 5.7) finds the same positive effect of *airborne-atopy* on *URI* but also that it lead to *severe-* but not *mild- LRI*. Figure 5.7 also indicates that *wheeze* contributed to *severe-LRI* in a subset of cases. The same network inferred after mean-imputing missing data, figure H.1, was essentially the same except that the negative effect on *URI* was missing and the edge from *wheeze* to *severe-LRI* was darker.

Dividing *severe-LRI* further into its *wheezy* and *febrile* subsets, shown in figure 5.8, we found the positive effect of *airborne-atopy* is restricted to *wheezy-LRI*, with a dark edge indicating near-unit posterior. It follows that *airborne-atopy* led to *wheezy-LRI*. The network inferred by mean-imputing missing data is again much the same, but with a darker edge from *wheeze* to *wheezy-LRI*, and the dashed edge to *URI* vanishes altogether.

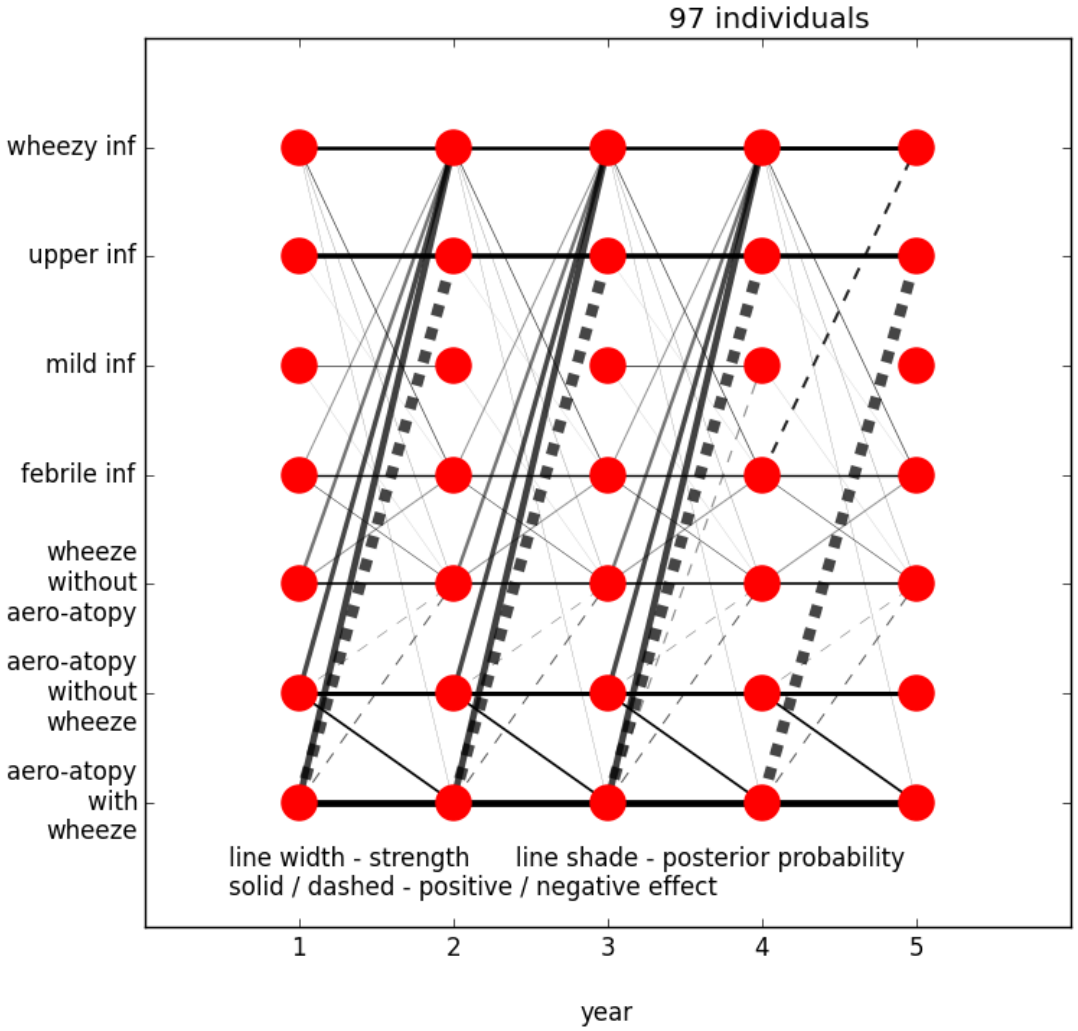


Figure 5.10: ***Airborne-atopy* with *wheeze* led against *URI* and to *wheezy-LRI*:** *Airborne-atopy* and *wheeze* also led to *wheezy-LRI* in isolation, but with weaker coefficients and lower posteriors, with the former being stronger and more probable than the later.

5.4.2 Atopy *and* wheeze *led against* URI

To better understand the apparent effect on *URI* we sought a relevant edge with near-unity posterior. To see if only certain subsets of *airborne-atopy* cases led against *URI*, we replaced *airborne-atopy* and *wheeze* in figure 5.7 and instead considered cases of both (*airborne-atopy* with *wheeze*), or just one (*airborne-atopy* only or *wheeze* only), as shown in figure 5.9. This immediately yielded a high posterior edge leading against

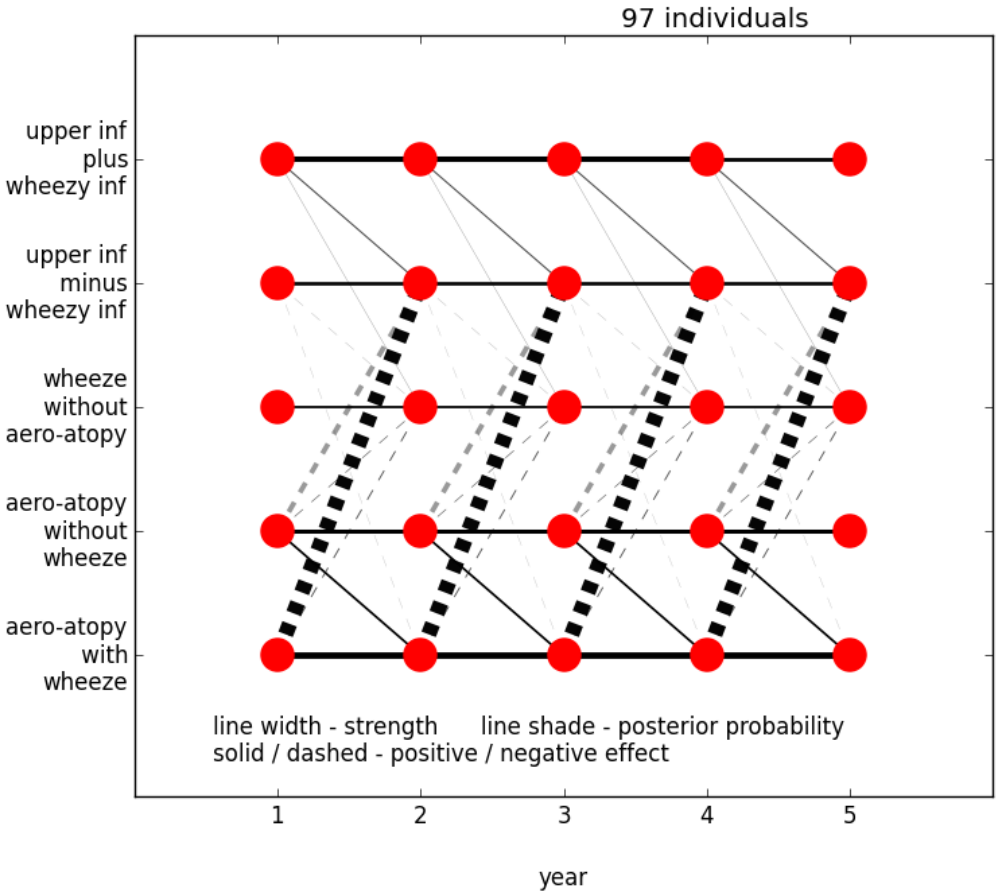


Figure 5.11: ***Airborne-atopy* with *wheeze*, and to a lesser extent without *wheeze*, led to *URI* becoming *wheezy-LRI*:** We verified this result with the χ-squared test. The p-values are presented in table 5.3.

URI, indicating that *aeroatopic-wheeze* led against *URI*. Repeating our above division of *severe-LRI* into *febrile-* and *wheezy- LRI*, shown in figure 5.10, found that the positive effects of *airborne-atopy* and *wheeze* were limited to *wheezy-LRI* with an even (slightly) higher posterior while the negative effect on *URI* by *aeroatopic-wheeze* was essentially unchanged. (The mean-imputed networks, figures 5.9 and 5.10 were essentially the same. Consistent with earlier mean-imputed networks, the differences were darker edges leading to *severe-LRI* and *wheezy-LRI*, respectively, and fainter edges leading against *URI*.) We are unable to comment on whether this is due to some interaction between *airborne-atopy* and *wheeze*, or if the presence of *wheeze* is an indicator of the severity

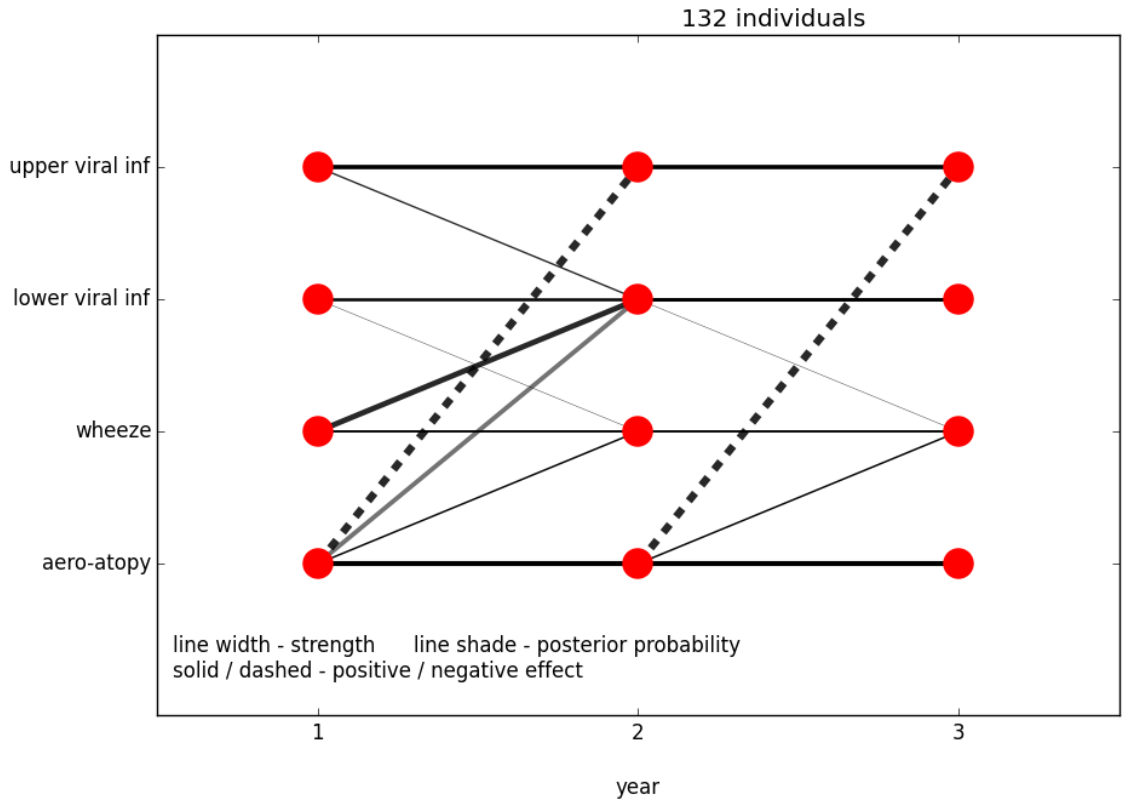


Figure 5.12: ***Airborne-atopy* led against *viral-URI*. In the first *year-of-life*, with low-to-intermediate posterior, it led to *viral-LRI*:** *Airborne-atopy* also lead to *wheeze* and all variables are ongoing.

of *airborne-atopy*.

5.4.3 Aeroatopic-wheeze *turned* URI *into* LRI

One would not expect *airborne-atopy* or *wheeze* to protect against infection, and the most biologically plausible interpretation of *aeroatopic-wheeze* leading against *URI* is that it makes it easier for infectious agents to enter the lungs, effectively allowing what would have been *URI*s to become *LRI*s. We have just seen that it is *wheezy-LRI*s in particular that are increased, so the effect of *aeroatopic-wheeze* on *URI*s would be to turn them into *wheezy-LRI*s.

A DBN to better test this statement would seek not the effect of *aeroatopic-wheeze* on *URI* and *wheezy-LRI* individually, but on the balance between them. One of the former becoming one of the latter is mathematically equivalent to taking a weight from

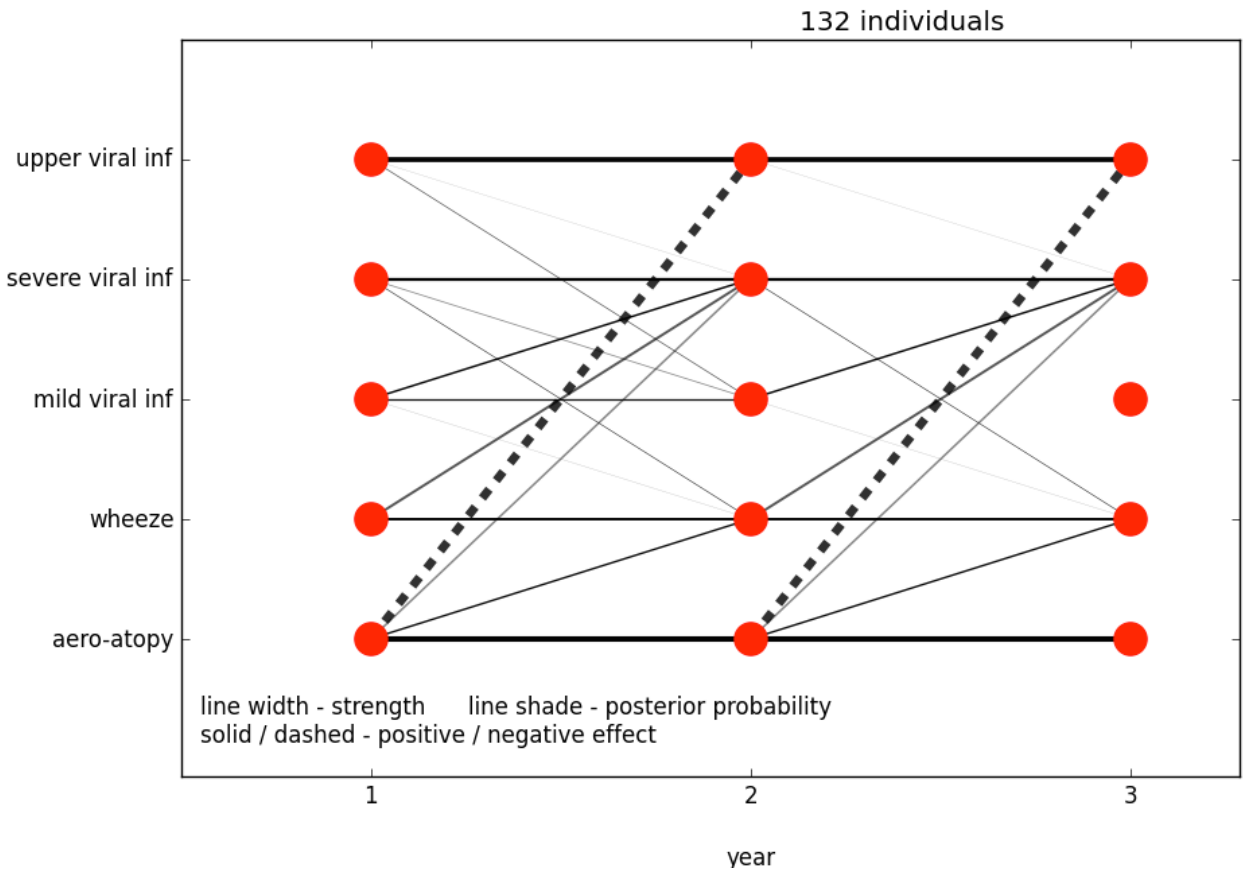


Figure 5.13: ***Airborne-atopy* led against *viral-URI* with intermediate posterior:**

one side of a balance and placing it on the other. The balance then moves in response to the change in the difference between the weight on either side, although the total weight on the balance scale is obviously the same. Analogously, it is the change in the difference between *URI* and *wheezy-LRI* while the sum total of these two infection types remains constant that corresponds to our proposed effect of *aeroatopic-wheeze* on *URI*. The corresponding network from imputed data was almost identical except that the negative edge leading to the difference between *URI* and *wheezy-LRI* is absent (see figure H.5).

Applying the χ-squared test to figure 5.11 further supported the result that *aeroatopic-wheeze* led to *wheezy-LRI* in the place of *URI*. The relevant p-values are shown in tables 5.3. The apparent effect is most significant within each *year-of-life*, indicating a

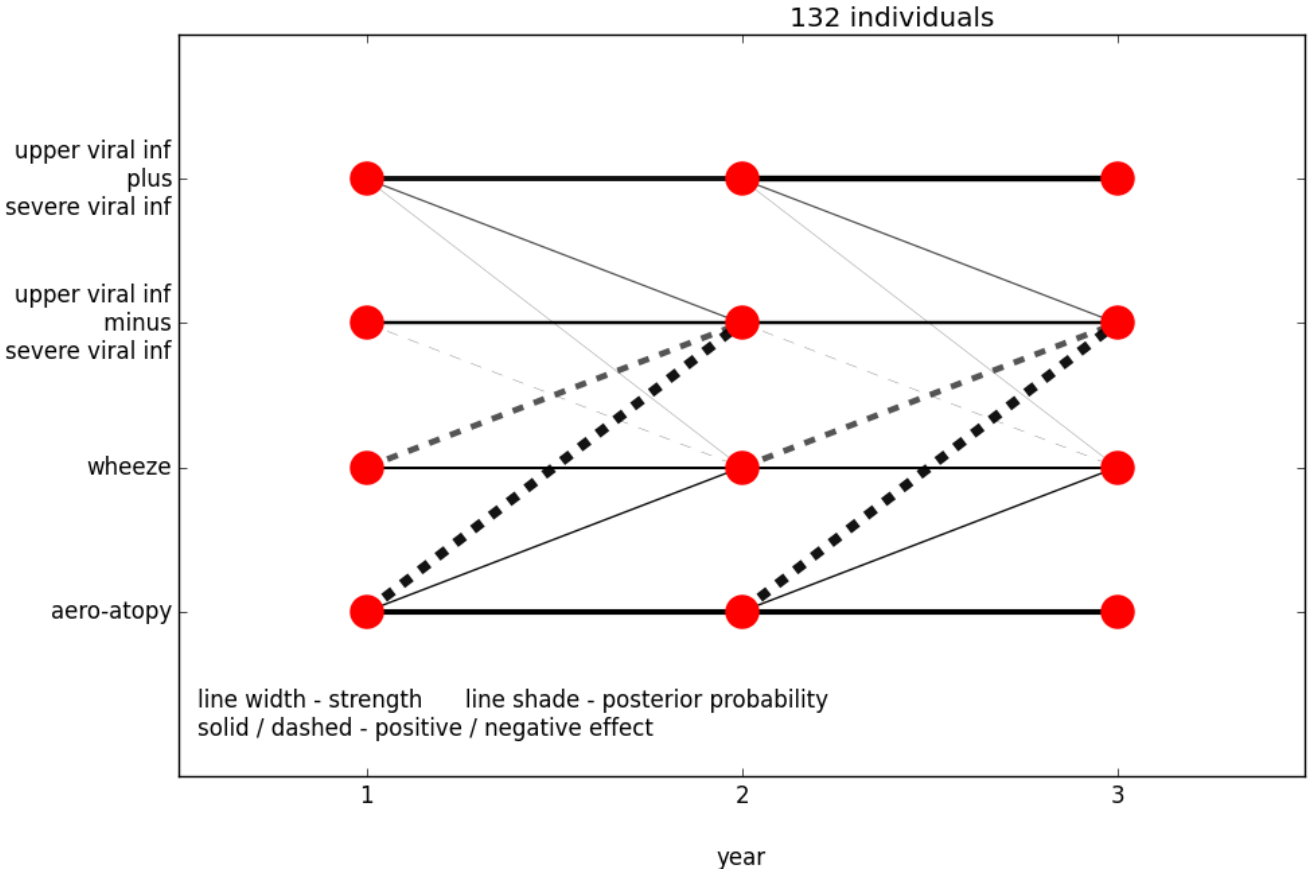


Figure 5.14: **Both *airborne-atopy* and *wheeze* led to an increase in *severe-viral-LRI* with a simultaneous decrease in *viral-URI*, the latter with only intermediate posterior:** A natural interpretation is *airborne-atopy* and *wheeze* allow what would have otherwise been a *viral-URI* to enter the lungs and become a *severe-viral-LRI*. We tested this link with the χ-squared test. The corresponding p-values are in table 5.4 and support the effect of *airborne-atopy* in this figure.

timescale significantly shorter than the yearly resolution of CAS data. In the first two *years-of-life* the p-values indicate an effect of earlier *aeroatopic-wheeze* on subsequent *URI* vs *LRI*, with no evidence of flow in the opposite direction. In later years the picture is less clear, although the most significant signal occurs within the given year.

By contrast, variation of the total number number of infections as a function of *aeroatopic-wheeze* never scored p-values less than 0.7.

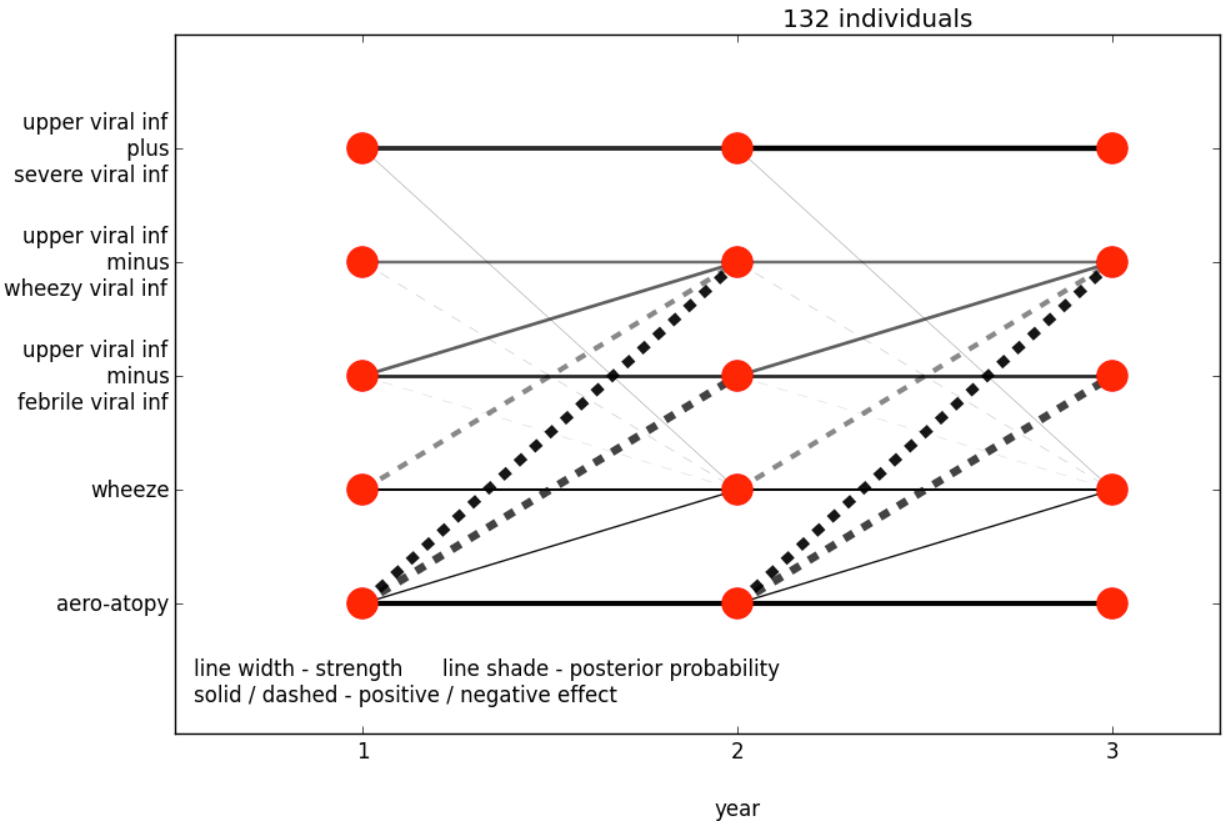


Figure 5.15: ***Airborne-atopy* led to an increase number of *wheezy-viral-LRI*s at the expense of *viral-URI*s. With intermediate posterior, both *airborne-atopy* and *wheeze* also led to an exchange of *viral-URI*s for *febrile-viral-LRI*s:**

5.4.4 The importance of viral *status*

Restricting the infections in figure 5.4 to be viral yielded the network in figure 5.12. The edges leading to *viral-LRI* from *airborne-atopy* are very similar but those from *wheeze* are clearly darker than their counterparts in figure 5.4. Edges leading against *viral-URI* from *airborne-atopy* are also darker than in figure 5.4.

Given the results of the previous subsection, we divided *viral-LRI* into *severe-* and *mild-* cases, which is shown in figure 5.13. The negative effect against *viral-URI* remains but the edges leading to *severe-viral-LRI* and *mild-viral-LRI* are faint and non-existent, respectively. There is further degradation when *severe-viral-LRI* is further separated into *wheezy* and *febrile* infections (not shown). This is surprising, given our discussion of figures 5.4 and 5.13.

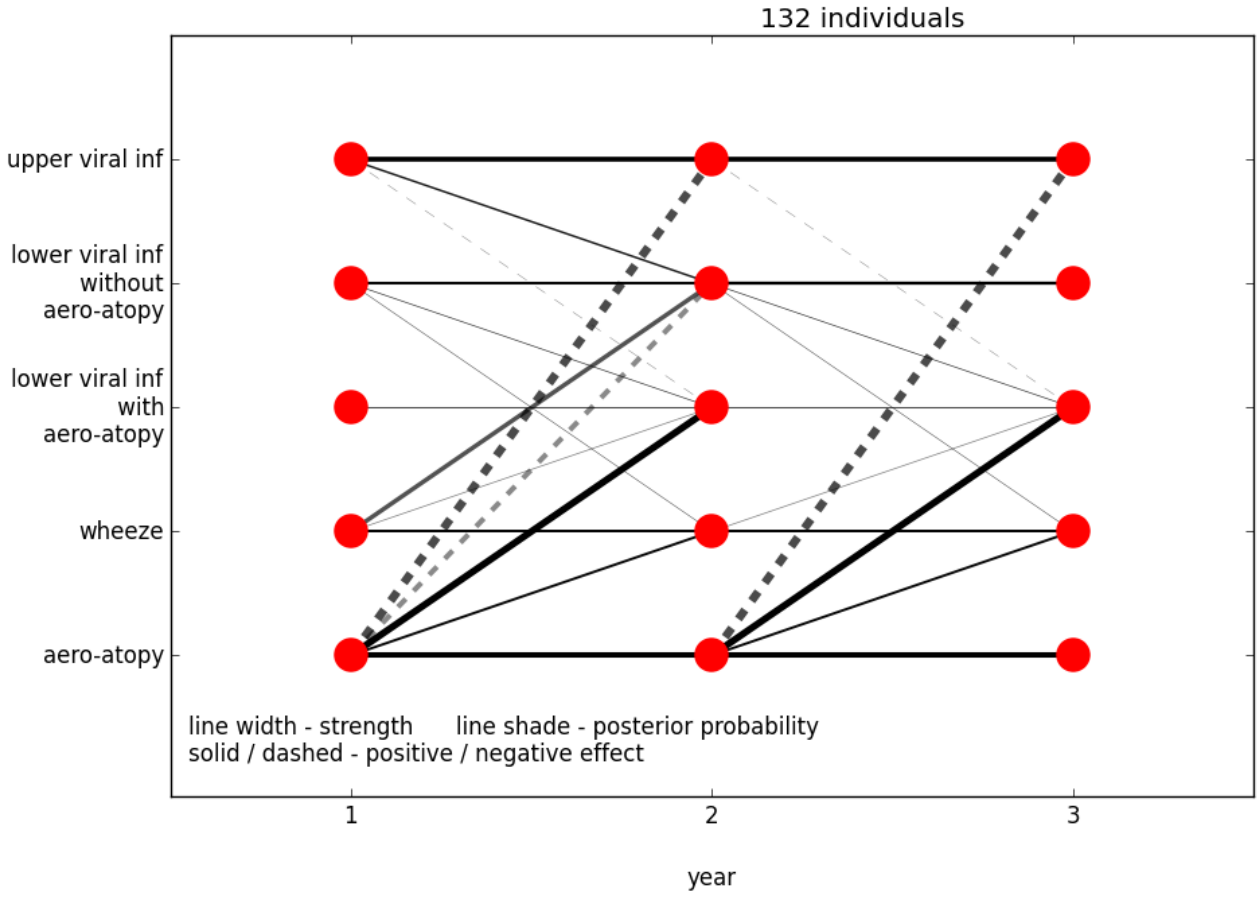


Figure 5.16: ***Airborne-atopy* led to *viral-LRI*s and not the other way around:**

Another effect which does not seem to be conserved by the restriction to *viral* infections is the effect of combining *wheeze* and *airborne-atopy*, like in figures 5.9 and 5.10. Indeed, the network obtained by making this variable change to figure I.1 found the corresponding link to *wheezy-viral-LRI* with a much lower posterior while in those based on the figures 5.12 and 5.13 the edge leading against *viral-URI* doesn't appear at all (not shown). We recall that *ARTIVA* samples across timesteps and infers phases, and in the first *year-of-life* there were only eight cases of *aeroatopic-wheeze*(see table C.2), with the condition more prevalent in the excluded later years. This makes it likely that there is insufficient statistical power to infer these edges for *viral-LRI*s, for which there are only three years of data. If that is the case then we might still be able to replicate the effect shown in figure 5.11 in a network restricted to viral infections.

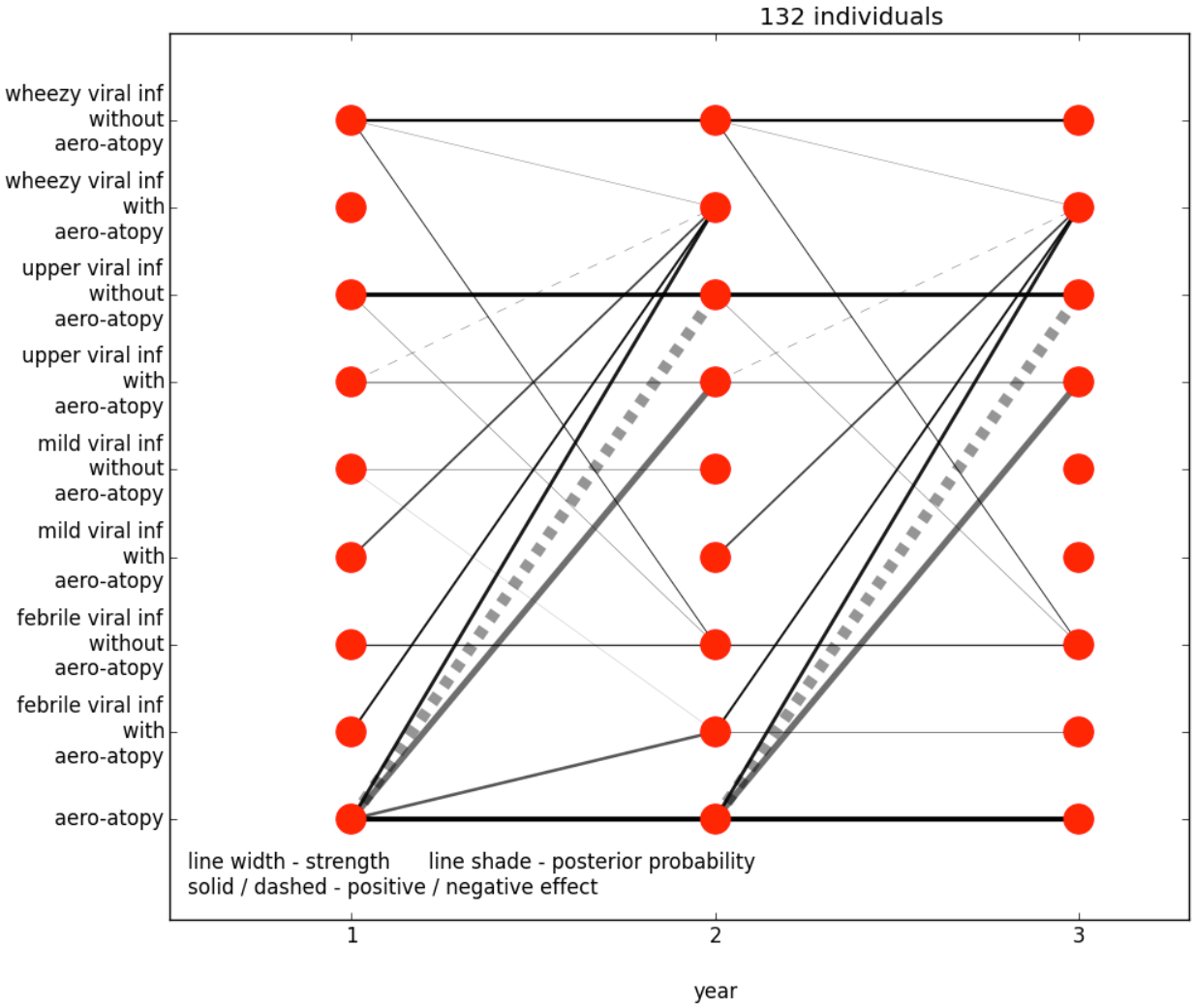


Figure 5.17: ***Airborne-atopy* led to *viral-LRI*, and not the other way around:** *Viral-LRI*s not accompanied by *airborne-atopy* were highly unlikely to lead to other *viral-LRI*s.

We therefore inferred networks to test if *airborne-atopy* and *wheeze* led to *viral* infections being *wheezy-LRI*s rather than *URI*s, shown in figures 5.14 and 5.15. We found a corresponding negative link from *airborne-atopy* to the difference between *viral-URI* and *severe-viral-LRI* in the former of these networks. In the latter we found a similar edge for *wheezy-viral-LRI* and another, though only of intermediate posterior, for *febrile-viral-LRI*.

The corresponding networks for *non-viral* infections (not shown), found no connec-

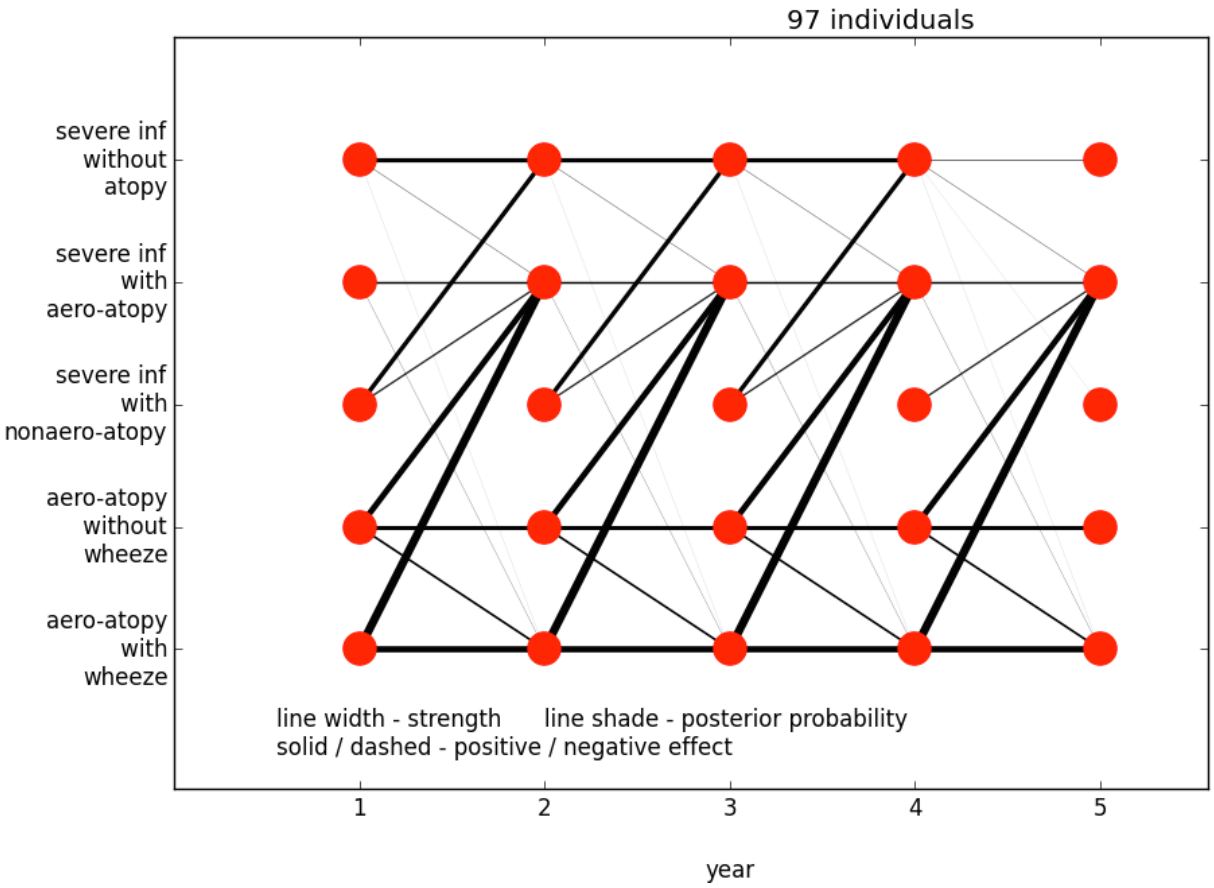


Figure 5.18: ***Airborne-atopy* led to *severe-LRI* and not the other way around:** *Nonairborne-atopy* typically resolved the following year.

tions with *wheeze* or *airborne-atopy*, but given the low numbers of *non-viral* infections, we refrain from drawing any conclusions.

5.5 *Viral-LRI*s and continuance of atopy

5.5.1 Viral-LRI*s accompanied by* airborne-atopy *did not lead to* airborne-atopy

Having seen that *airborne-atopy* led to *viral-LRI* and not the other way around, we considered whether there might be a positive feedback loop, in which infection accompanied by *airborne-atopy* led to aggravated *airborne-atopy*. We focussed on *airborne-atopy* since *nonairborne-atopy* was found to not contribute to later *wheeze* as per our discussion of figures 5.1 through 5.4. The reader is reminded again of the legend in

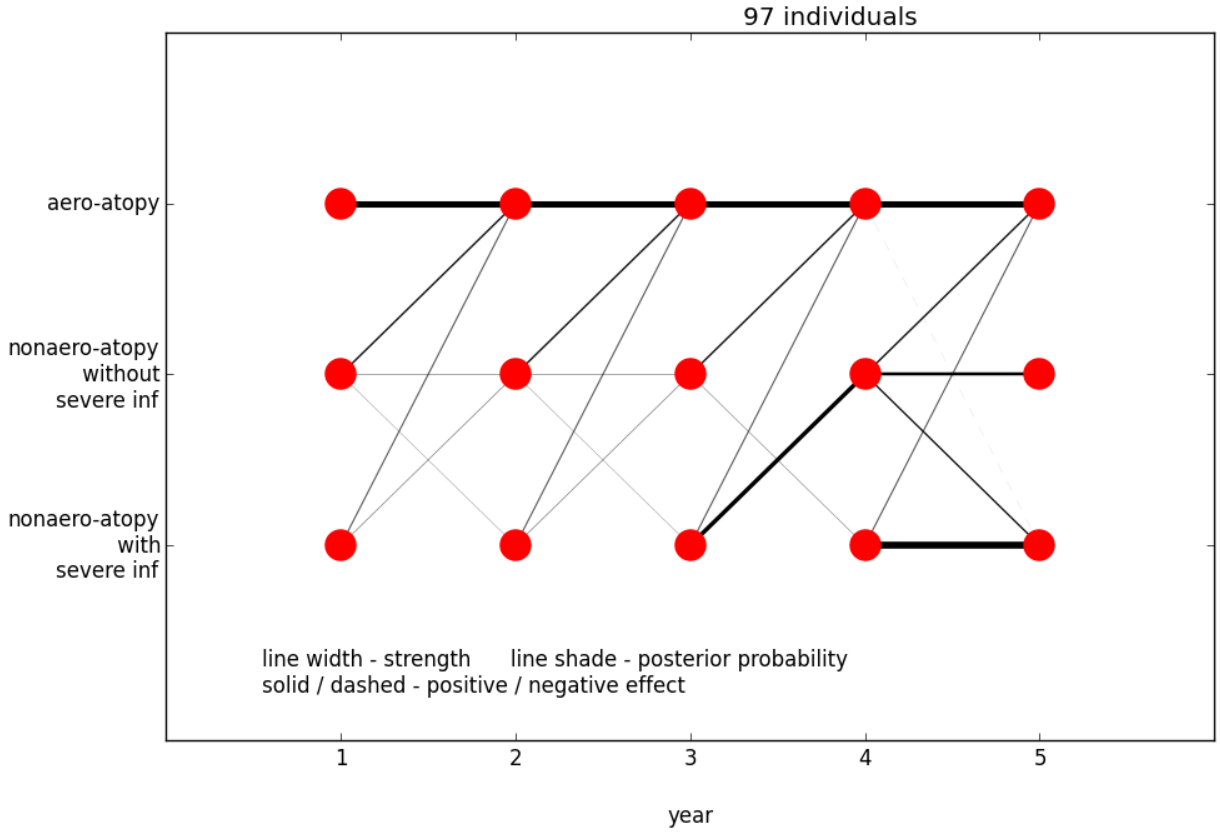


Figure 5.19: **There is no evidence that *severe-LRI* led to *airborne-atopy* if it wasn't already present, even in the presence of *nonairborne-atopy*:**

appendix G.

In figures 5.16 and 5.17 we considered whether the *viral-LRI*s were accompanied by *airborne-atopy*. If *severe-*, *febrile-* or *wheezy- viral-LRI* contributed to the persistence of *airborne-atopy* then the signal would be stronger from those infections for which *airborne-atopy* was already present. After all, *airborne-atopy* must be present if it is to persist! What these figures indicate however is that *airborne-atopy* still led to *viral-LRI*and not the other way around.

Figure 5.18 also indicates that *airborne-atopy*, with or without *wheeze*, led to *severe-LRI* and not the other way around.

There has been discussion in the literature [6, 16] about whether *viral* infection can lead to *atopic* responses in the lungs. All DBNs shown so far indicate that any

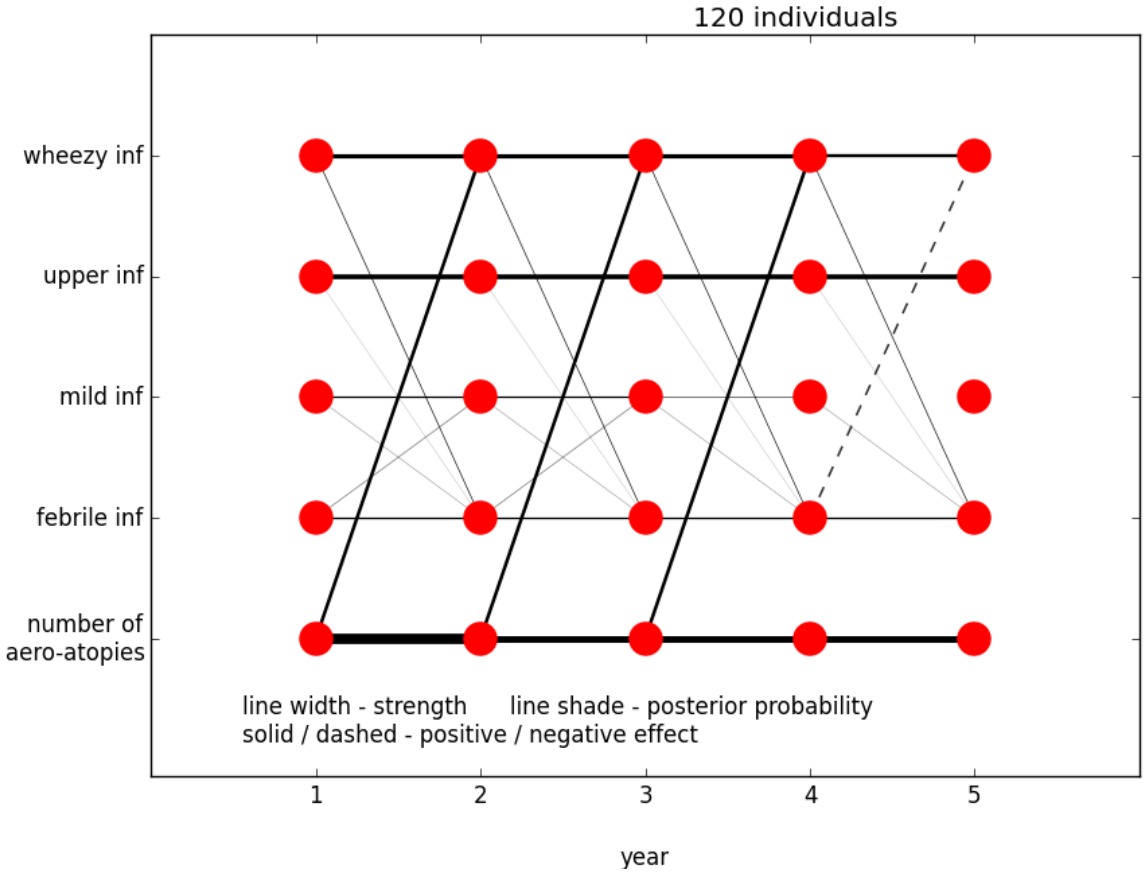


Figure 5.20: **The number of aeroatopies, and not infections, drives the number of aeroatopies:** Multiple aeroatopies led to *wheezy-LRI*, but there was negligible effect on other infection types.

causal influence is in the opposite direction, but we considered whether *nonairborne-atopy* might be more likely to lead to *airborne-atopy* given the co-occurence of *severe-LRI*. Figure 5.19 clearly indicates that the link is tenuous at best, with the posterior probability being lower when *severe-LRI* is present.

We therefore continue to find that it is *airborne-atopy* that led to *LRI* and not the other way around, although the figure 5.18 seems to indicate that the copresence of *severe-LRI* with *airborne-atopy* led to ongoing *airborne-atopy*. We cannot yet comment on whether this is causal or due to the *atopic* being infection-prone.

	febrile-LRI		
aeroatopy-number	previous year	same year	following year
First *year-of-life*	NA	$\mathbf{< 2.2 \times 10^{-16}}$	$\mathbf{2.22 \times 10^{-7}}$
Second *year-of-life*	1.62×10^{-12}	$\mathbf{3.65 \times 10^{-14}}$	1.82×10^{-4}
Third *year-of-life*	$\mathbf{1.88 \times 10^{-12}}$	$\mathbf{1.19 \times 10^{-13}}$	1.61×10^{-6}
Fourth *year-of-life*	$\mathbf{2.12 \times 10^{-15}}$	2.51×10^{-12}	7.2×10^{-5}
Fifth *year-of-life*	$\mathbf{2.51 \times 10^{-3}}$	6.55×10^{-3}	NA

Table 5.5: χ-squared test p-values of the interaction between *aeroatopy-number* and the number of *febrile-LRI* in the same and adjacent *years-of-life*, where the *febrile-LRI* are accompanied by *airborne-atopy*. Figure 5.20 does not show the relationship suggested here, but figure 5.21 reflects it quite closely, once the restriction to *purely-febrile-LRI* is made and the copresence of of *airborne-atopy* is required.

5.5.2 Severe-LRI*s and the number of* atopic *triggers*

In subsection 3.2.2 we found *atopy-number* in the first *year-of-life* to be an excellent predictor (AUC = .81) of *atopic-wheeze* in the fifth *year-of-life*. It appears to have little interaction between the number of infections, with only a mild effect on *wheezy-LRI* seen in figure 5.20. The situation changed if infections were divided according to whether they were accompanied by *airborne-atopy*, (figure 5.21). In common with subsequent figures, figure 5.21 finds a link from *aeroatopy-number* to *febrile-* and *wheezy- LRI* in the first *year-of-life* but the edges run from *febrile-LRI* to *aeroatopy-number* from the second *year-of-life* onwards. The face-value interpretation of this observation is that *febrile-LRI* led to an increasing, or prevented a decreasing, number of *atopic* triggers in those that were *aeroatopic* already. However it is also plausible that a propensity for increasing *atopy-number* coincides with susceptibility to *febrile-LRI*, so we cannot state on this evidence that *febrile-LRI*s aggravate *atopy*.

5.5.3 Effect of infection on aeroatopy-number

Figure 5.22 indicates that in the first *year-of-life* it is *aeroatopy-number* that led to *viral* infection and not the other way around. It further indicates that the only infections to link with *aeroatopy-number*, or even other infections for that matter, were accompanied by *airborne-atopy*.

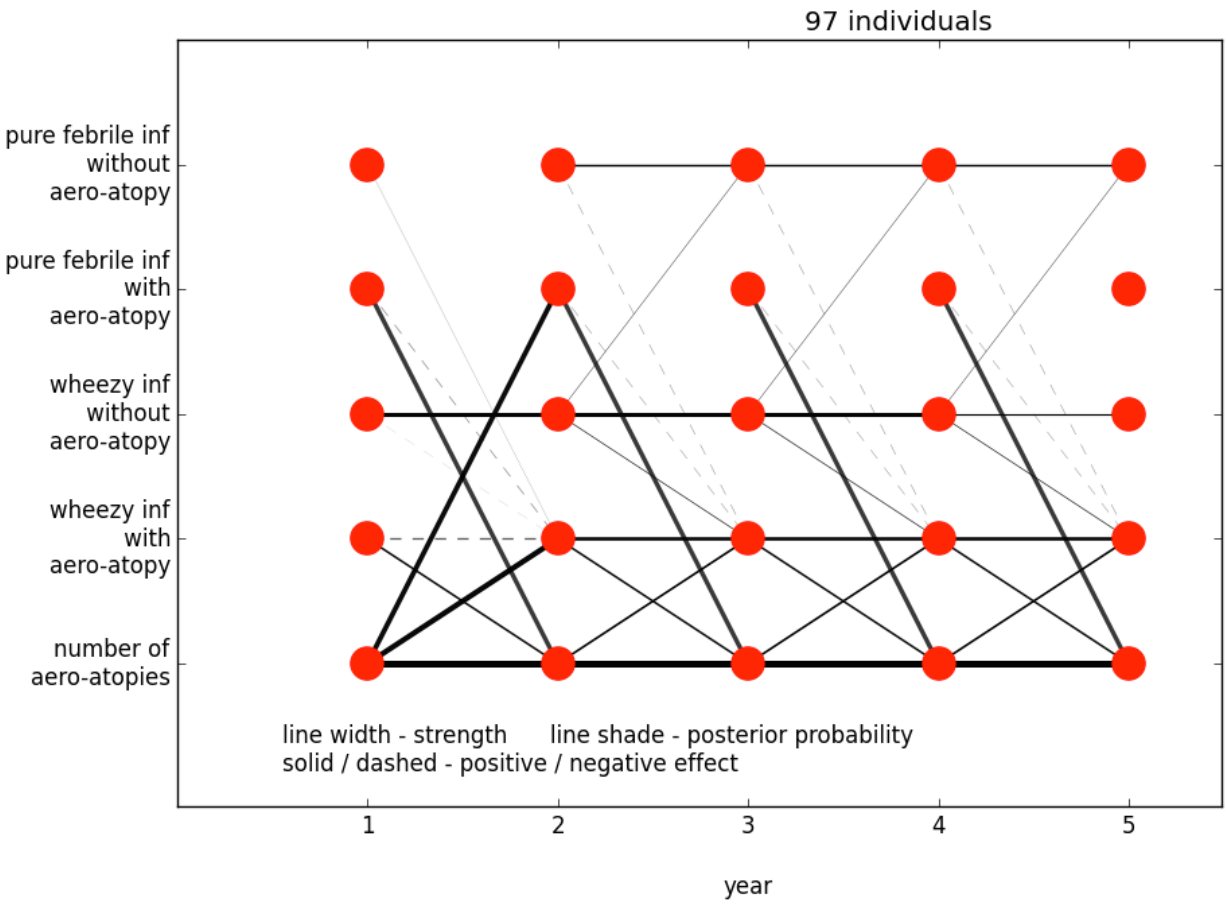


Figure 5.21: **All infections are assumed to be accompanied by *airborne-atopy*. *Purely-febrile-LRI* led to *aeroatopy-number*. *Aeroatopy-number* led to *wheezy-LRI*, especially in the first *year-of-life*, and led to *febrile-LRI* in the first *year-of-life*:**

5.5.4 Causality runs from aeroatopy-number *to* severe-viral-LRI *in the first* year-of-life

Figure 5.23 indicates that *mild-viral-LRI* accompanied by *airborne-atopy* led against lower *aeroatopy-number*.

It is biologically implausible that *mild-viral-LRI* would actively lower *aeroatopy-number*, and combining this with the failure of *severe-viral-LRI* to affect *aeroatopy-number* in figures 5.22, we found that that the infections' lack of severity indicates that the *aeroatopy-number* is not increasing, *i.e.* a propensity to increasing *aeroatopy-number* is associated with *severe-viral-LRI* among those with *airborne-atopy*. Indeed,

	wheezy-LRI		
aeroatopy-number	previous year	same year	following year
First *year-of-life*	NA	$\mathbf{1.70 \times 10^{-14}}$	$\mathbf{3.18 \times 10^{-9}}$
Second *year-of-life*	8.24×10^{-14}	1.77×10^{-12}	$\mathbf{4.76 \times 10^{-16}}$
Third *year-of-life*	6.68×10^{-9}	$< \mathbf{2.2 \times 10^{-16}}$	$\mathbf{1.61 \times 10^{-6}}$
Fourth *year-of-life*	3.77×10^{-9}	$\mathbf{4.55 \times 10^{-13}}$	$\mathbf{1.62 \times 10^{-5}}$
Fifth *year-of-life*	$\mathbf{5.22 \times 10^{-9}}$	1.33×10^{-7}	NA

Table 5.6: χ-squared test p-values of the interaction between *aeroatopy-number* and the number of *wheezy-LRI* in the same and adjacent *years-of-life*, where the *wheezy-LRI* are accompanied by *airborne-atopy*. For the first four *years-of-life* there is strong evidence of an effect, but without a clear chronological direction since it takes place on a timescale much shorter than one year. There is agreement with the network in figure 5.20 that the chronological arrow goes from *aeroatopy-number* to *wheezy-LRI* in the first *year-of-life*, and the opposite direction in the fifth, but they disagree as to when that reversal takes place. However both approaches, each in their own way, indicate a corresponding uncertainty in their edges.

	mild-viral-LRI		
aeroatopy-number	previous year	same year	following year
First *year-of-life*	NA	$\mathbf{2.166 \times 10^{-6}}$	2.292×10^{-4}
Second *year-of-life*	**.235**	$< \mathbf{2.2 \times 10^{-16}}$	.1785
Third *year-of-life*	$\mathbf{1.855 \times 10^{-7}}$	3.742×10^{-6}	NA

Table 5.7: χ-squared test p-values of the interaction between *aeroatopy-number* and the number of *mild-viral-LRI* in the previous, same and following year, where the *mild-viral-LRI* are accompanied by *airborne-atopy*. While there is no support for the specific edge connecting first-year *mild-viral-LRI* to second-year *aeroatopy-number*, there is strong support for an effect acting within each *year-of-life*, especially second year. This supports our interpretation of figure 5.23 that susceptibility to *severe-viral-LRI* accompanying *airborne-atopy* indicates a predisposition for increasing *aeroatopy-number*.

we have argued in section 5.4.4 that *airborne-atopy* led to *severe-viral-LRI*, especially in the presence of *wheeze*. This is consistent with the aforementioned analysis by Wennergren and Kristjánsson [60] finding that *severe RSV* infection followed by wheeze indicated an underlying mutual cause. This result is important so we verified it with the χ-squared test, shown in table 5.7. This demonstrated a functional relationship between *aeroatopy-number* and *mild-viral-LRI* in the presence of *airborne-atopy*, with a very significant signal in the second *year-of-life*. As an additional check, we tested for a corresponding relationship with *mild-viral-LRI* in the absence of *airborne-atopy*. The required negative result is presented in table I.1.

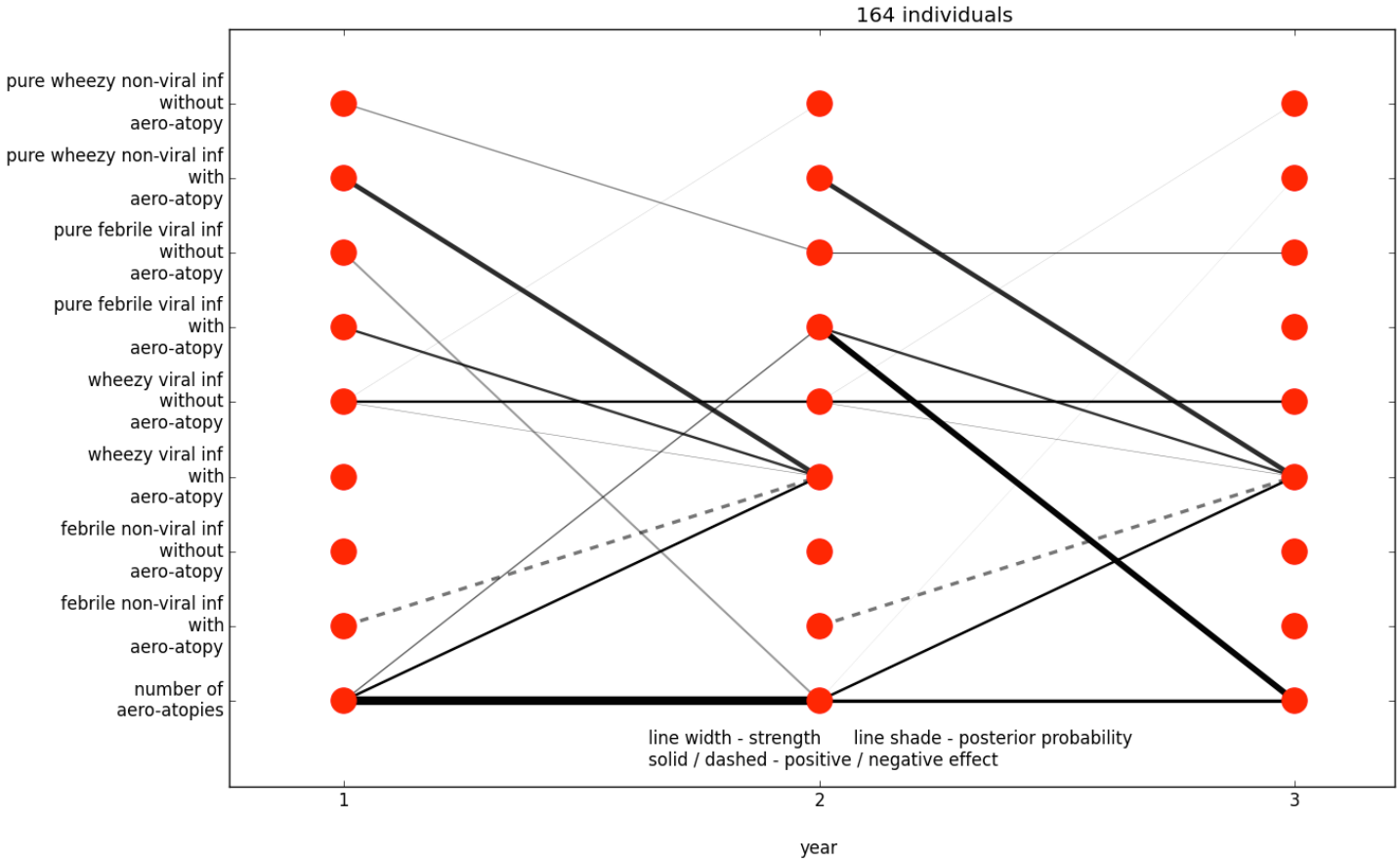


Figure 5.22: **Nothing in the first *year-of-life* led to *aeroatopy-number* except itself, but *aeroatopy-number* in the first two *years-of-life* led to *wheezy-viral-LRI* (with *airborne-atopy*). *Febrile-viral-LRI* with *airborne-atopy* in the second *year-of-life* led to *aeroatopy-number*. Infections not accompanied by *airborne-atopy* did not lead to other infections, and only *wheezy-viral-LRI* is ongoing:**

To test the robustness of this result we split *severe-viral-LRI* into its *febrile* and *wheezy* subtypes, as shown in figure 5.24. Not only did first-year *aeroatopy-number* lead to *febrile-* and *wheezy- viral-LRI*, but first-year *mild-viral-LRI* led against *aeroatopy-number*, just as it did in figure 5.23. This supports our earlier finding that it is *airborne-atopy* that led to *severe-LRI* and not the other way around.

These figures also indicate that *severe-LRI*, and especially *febrile-LRI*, led to higher *aeroatopy-number* in those that were already *aeroatopic* from the second *year-of-life* onwards. Given that our earlier DBNs found that *airborne-atopy* led to *wheezy-LRI* but not *febrile-LRI*, regardless of *viral*-status, this would seem to be of significance.

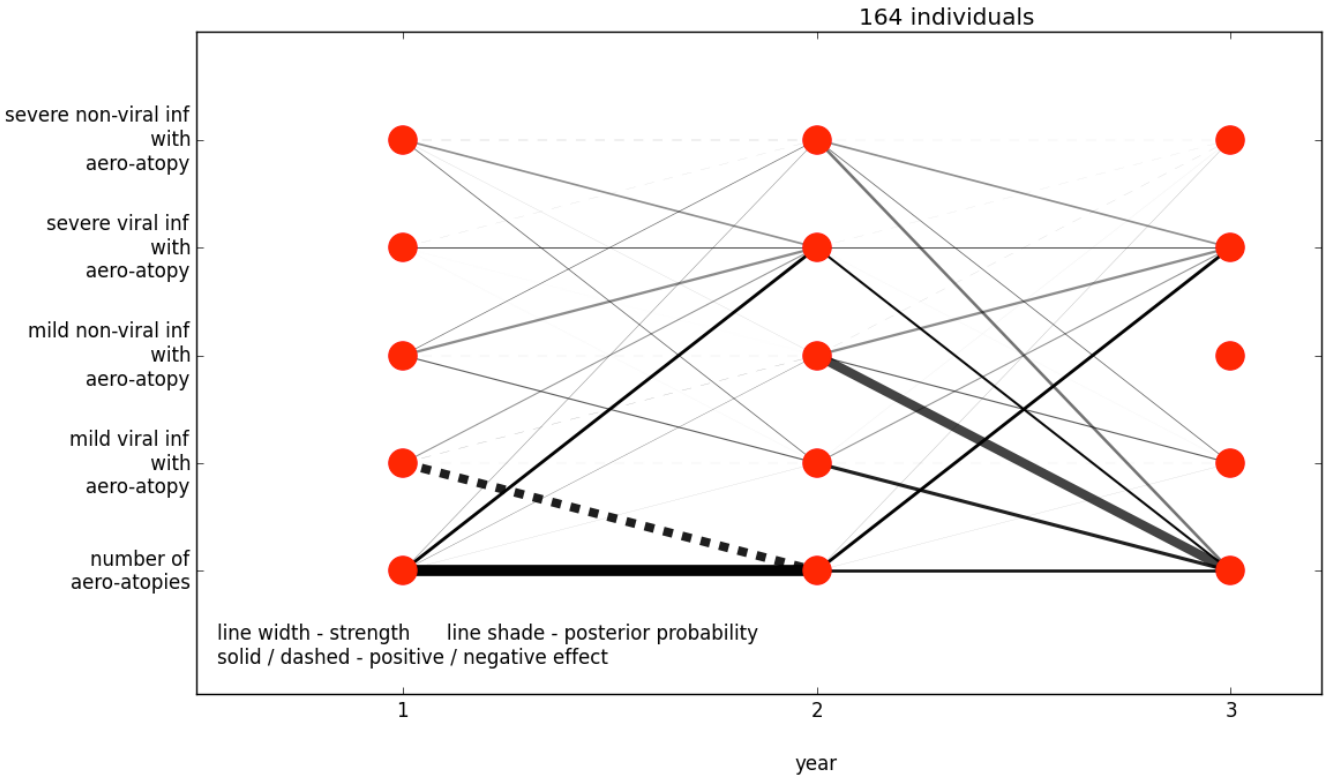


Figure 5.23: ***Mild-viral-LRI* led against *aeroatopy-number*, indicating that *aeroatopy-number* causes *severe-viral-LRI*, and not the other way around:**

Whether *febrile-LRI* was the actual causal agent, or was merely associated with it, cannot be determined here. The root causes of asthma are believed to occur within the first two *years-of-life* [5,173], but protection against *febrile-LRI* or *febrile-viral-LRI* may be an option for arresting pathogenesis.

Table I.1 checks the dependence on *airborne-atopy* of infections without *airborne-atopy*, and finds a greatly reduced significance as required by figures 5.23 and 5.24.

5.6 *IgE* dynamics and future *wheeze*-status

How *IgE* titres developed over time differed between those who went on to exhibit *wheeze* in the fifth *year-of-life* and those who did not. Our discussion of this will heavily involve the occurence of *IgE* titres blowing out to enormous values among *atopic*s. Like most of the variables already studied, our networks show that the value of a variable in a given time-step is dependent on its value in the previous time step. The important observable

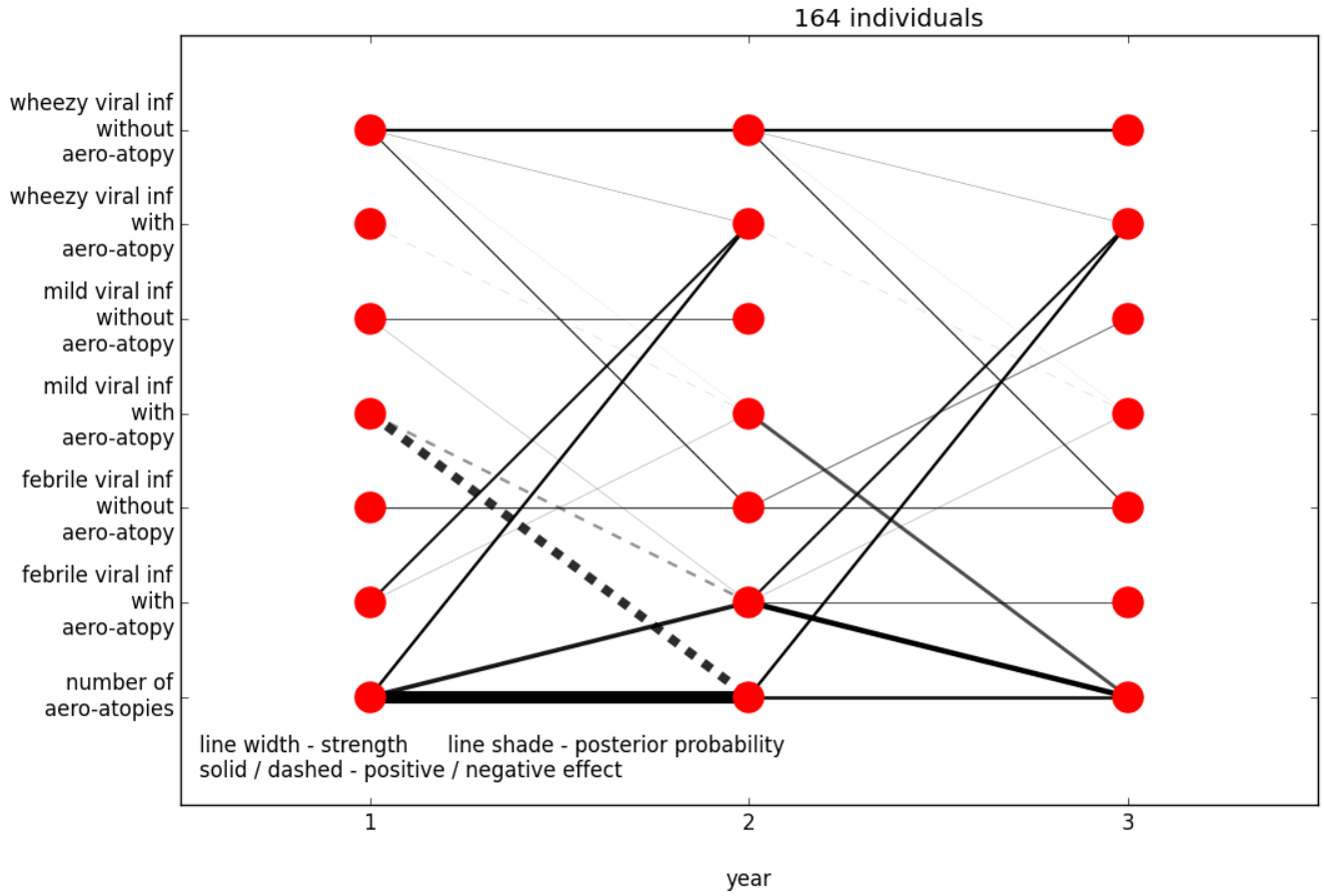


Figure 5.24: **First-year *mild-viral-LRI*, in the presence of *airborne-atopy*, led against *aeroatopy-number*. This apparent protection indicates that it is *airborne-atopy* that led to *severe-viral-LRI*, with failure to be *severe* indicating that *aeroatopy-number* was unlikely to increase. *Aeroatopy-number* in the first *year-of-life* led to both *febrile*- and wheezy- *viral-LRI* (with *airborne-atopy*):**

difference is that for *IgE*-related variables the dependence is often much stronger, so that instead of merely maintaining a condition the variables' value is growing exponentially. We sometimes say that the variable in question is *self-stimulating*.

Consider the networks shown in figures 5.25 and 5.26. In the network corresponding to fifth-year *wheeze*, there is a marked tendency for exponential growth of *IgE*. This is much less marked for the non-*wheeze* network, figure 5.25, where *IgE* increase, especially for *house-dust-mite IgE*, is much more dependent on the *IgE* of other allergens.

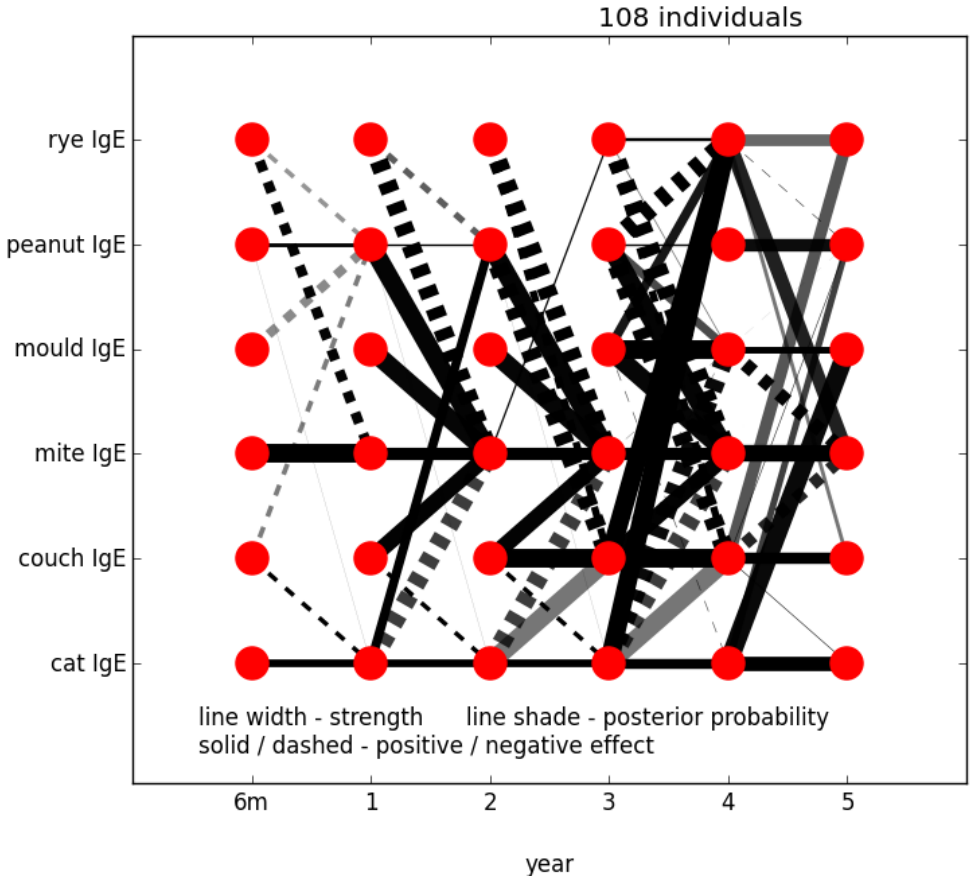


Figure 5.25: ***House-dust-mite IgE* showed exponential growth in the first *year-of-life*. *Couch-*, *mould-* and *peanut- IgE* led to *house-dust-mite IgE* from the first *year-of-life*, but not the first six *months-of-life*. *Rye-* and *cat- IgE* led against *house-dust-mite IgE* in the same period:** Network generated from participants who did not exhibit *wheeze* in their fifth *year-of-life*. Additional, later cross-linking among other *IgE*s.

Figures 5.27 and 5.28 show the corresponding networks with the logarithm of those variables. We see in these figures that in those who went on to exhibit fifth-year *wheeze* it was (log of) *house-dust-mite IgE* that was leading to other *IgE*s. It should not be a surprise that the *IgE* dynamics leading to fifth-year *wheeze* were manifest at a logarithmic scale. The *IgE* distributions of those who developed *wheeze* in the fifth *year-of-life* were very heavily skewed, even in comparison to those who did not.

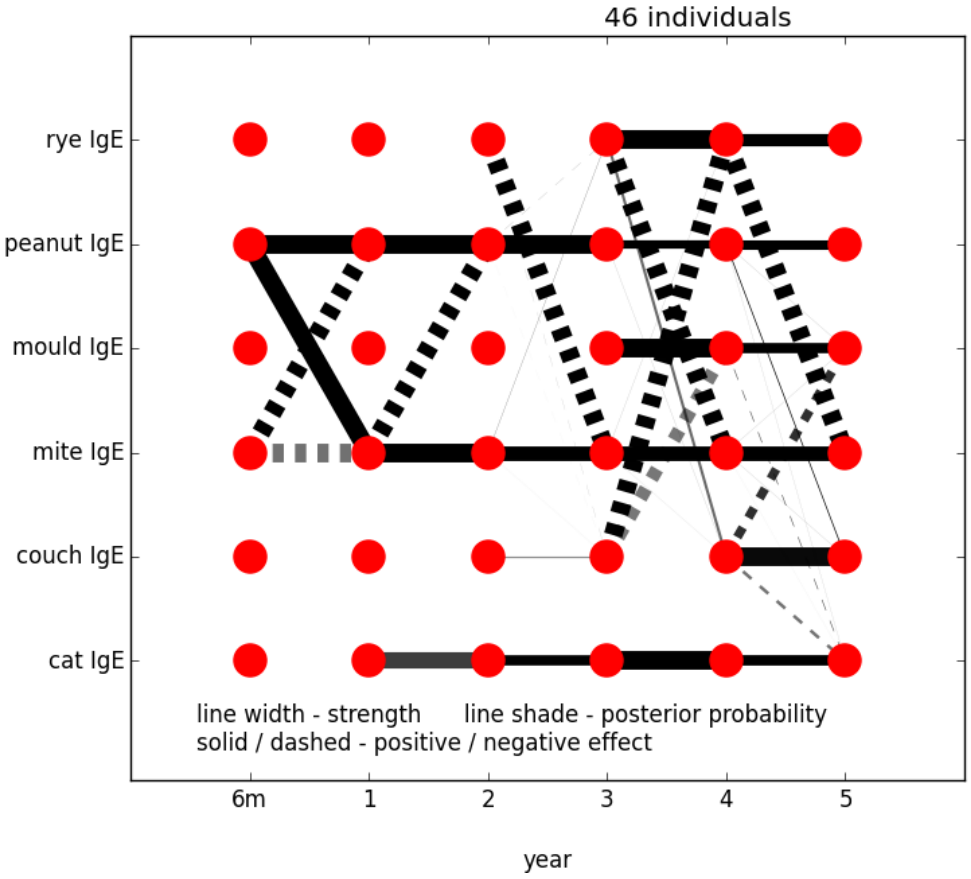


Figure 5.26: ***House-dust-mite IgE* lead against itself in the first six *months-of-life*, with intermediate posterior, and lead against *peanut-IgE* also in that and the next timestep. *Peanut-IgE* in the first six *months-of-life* lead to *house-dust-mite IgE*. All other cross-linking is negative, including *rye* and *couch*:**
Network generated from participants who exhibit *wheeze* in their fifth *year-of-life*.

5.7 Relations among atopic allergens

We make some observations about the dynamics of *IgE* titres.

Figure 5.26 indicates various relationships among the different *IgE*s. The first is a link from *peanut-IgE* in the first six months to *house-dust-mite IgE* in the first *year-of-life*. This explains how *wheeze* accompanied by *peanut atopy* in the fifth *year-of-life* is so well predicted by (log of) *house-dust-mite IgE* in the second *year-of-life*.

Another notable feature is the self-stimulated growth of *mould-IgE*, which begins

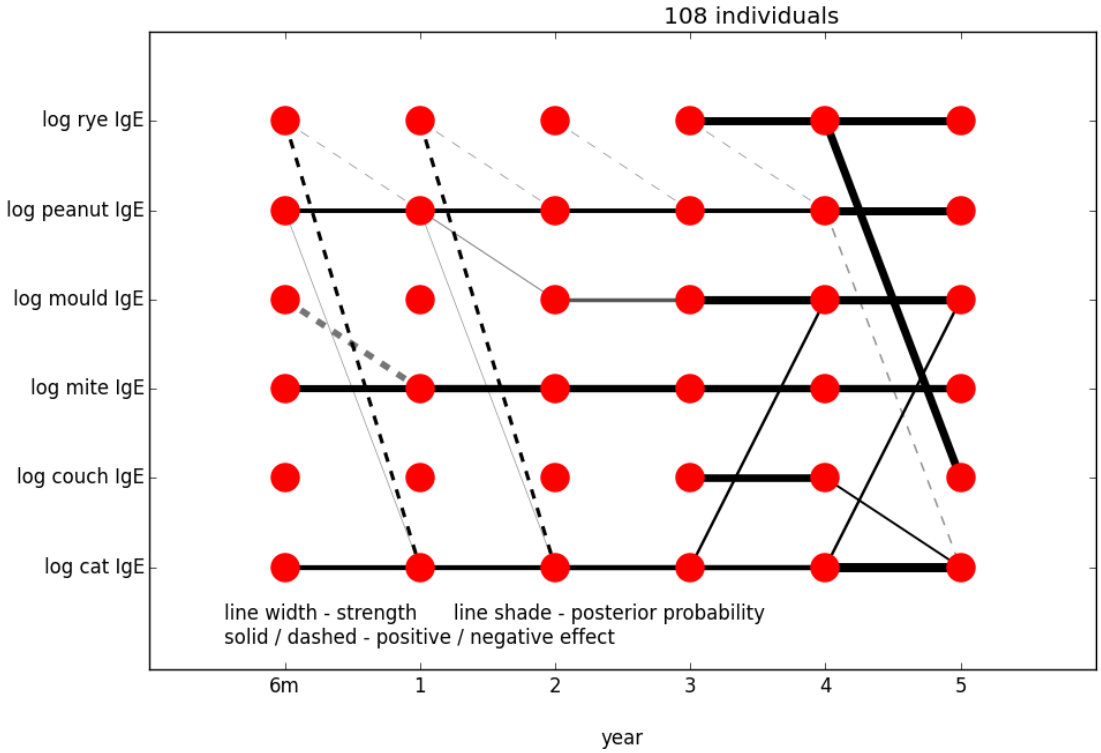


Figure 5.27: **Network among participants who did not *wheeze* in the fifth *year-of-life*. There is only minimal cross-linking between different *IgE* allergens at the logarithmic scale:**

in the third *year-of-life*. The reader may recall from section 3.4 that *mould*-specific *atopic-wheeze* was difficult to predict from second-year *IgE* but was well-predicted by *mould-IgE* in the third *year-of-life*.

5.7.1 Multiple atopies and IgE *titres*

Given these inter-allergen dependencies, it seemed prudent to consider the interaction of *atopy-number* with the *IgE* of various allergens. Figure 5.29 indicates that *atopy-number* led to (log of) *house-dust-mite IgE* in the first *year-of-life*, but otherwise it was (log of) *peanut-IgE* and mostly *house-dust-mite IgE* leading to *atopy-number*. In conjunction with the last four figures, especially figures 5.25 and 5.28, we see a pattern of *atopy* development, beginning with elevated *peanut-IgE* leading to elevated *house-dust-mite IgE* and then on to one or more other elevated *IgE*s. The mutual interaction

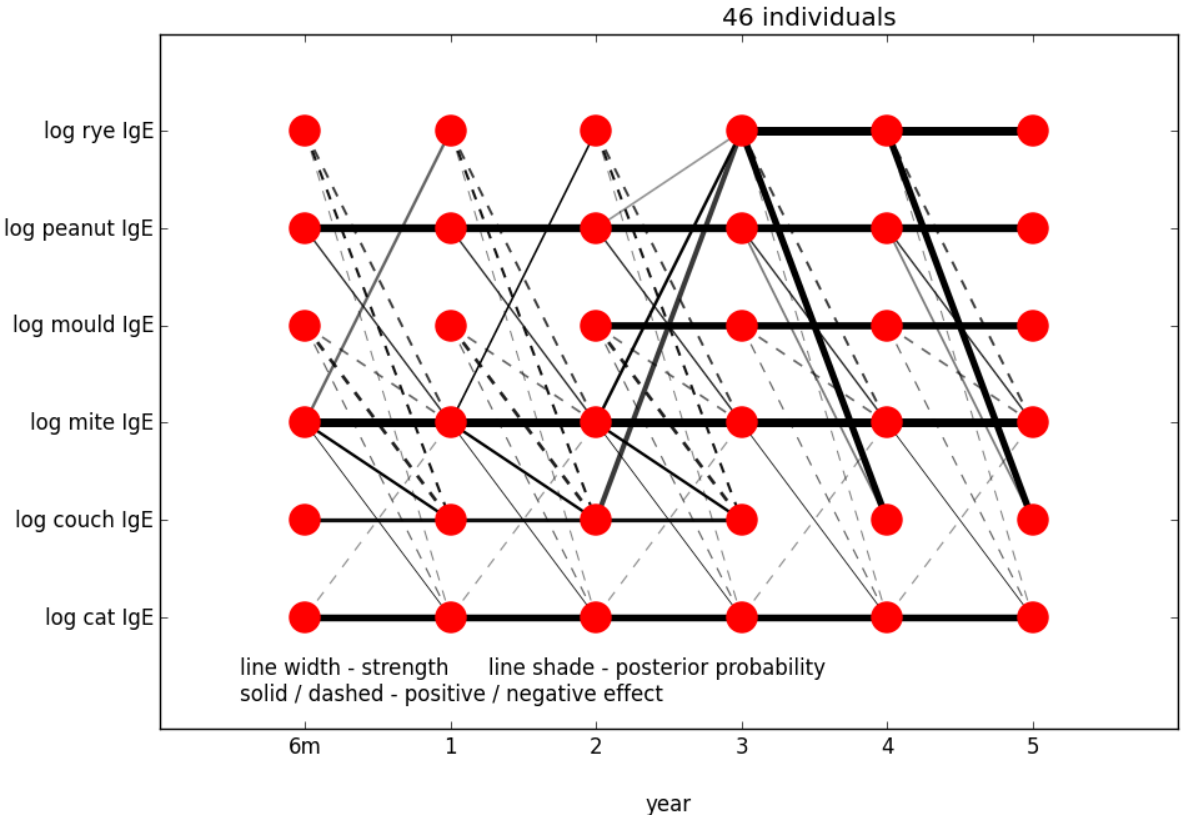


Figure 5.28: **Network from participants who exhibited *wheeze* in their fifth *year-of-life*. This network differs from figures to edges leading *from* (log of) *house-dust-mite IgE*:** There is also some later-year positive cross-linking between (log of) *rye-IgE* and (log of) *couch-IgE*, unsurprising since they are both grasses, and negative faint cross-linking in the first two *years-of-life*.

between *atopy-number* and (log of) *house-dust-mite IgE* is consistent with the multi-allergen endotype discussed in section 3.4. The significance of the other cross-linking among the different allergens is not clear.

Replacing *atopy-number* with *aeroatopy-number* (see appendix I.3) gives a similar graph, though with more faint edges among allergens and no link from first-year *atopy-number* to second year (log of) *infant-phadiatop-IgE*, and we prefer the "cleaner" graph of *atopy-number*.

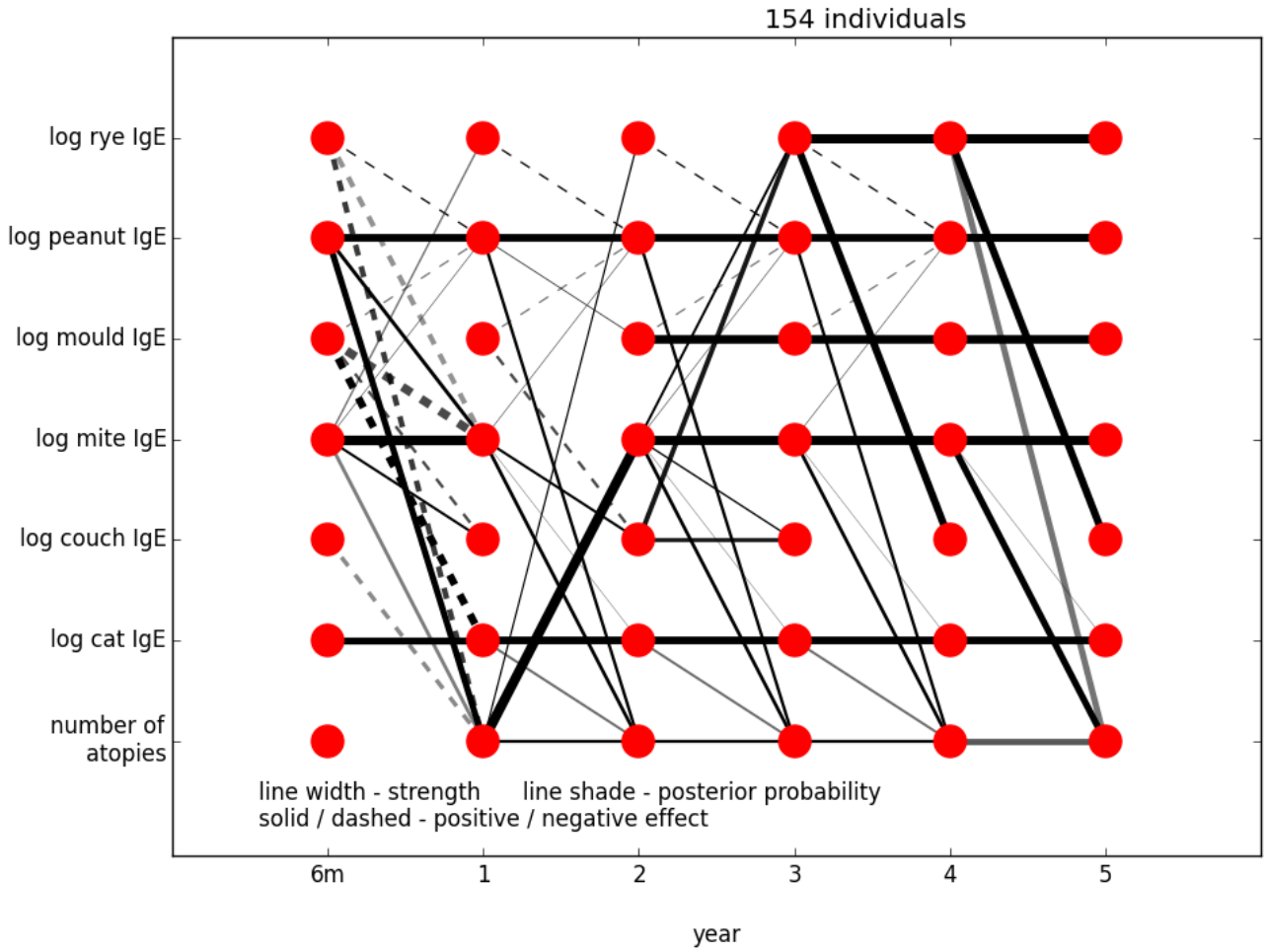


Figure 5.29: **(Log of) *peanut-IgE* and *house-dust-mite IgE* were the drivers of increasing *atopy-number*. *Atopy-number* in the first *year-of-life* lead to (log of) *house-dust-mite IgE*:**

5.8 Altered interleukin dynamics

We did not find any obvious links between interleukin and *wheeze*. There was too much missing data to find a DBN from those who did develop fifth-year *wheeze*, and we were further inhibited by lack of data from the third and fourth *years-of-life*.

We found interleukin networks for ovalbumin and for tetanus. However, clear differences were found among networks inferred from all participants and those inferred from those who went to exhibit, and not exhibit, *wheeze* in the fifth *year-of-life*.

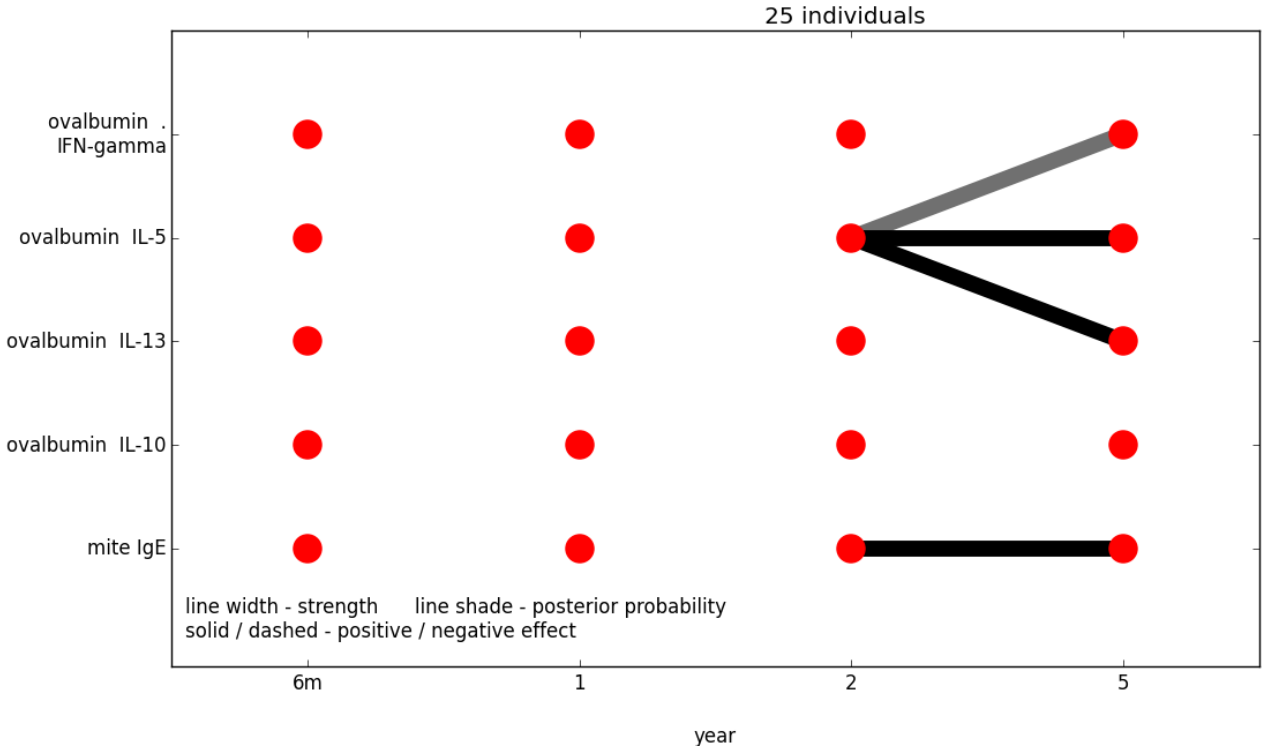


Figure 5.30: **Among those who did not exhibit fifth-year *wheeze*, ovalbumin IL-5 in the second *year-of-life* led to itself and ovalbumin IL-13 and IFN-γ, the latter with only intermediate posterior. The dynamics are very different from those of the unconditional network shown in figure 5.31:**

Ovalbumin IL-5 in the second *year-of-life* led to ovalbumin IFN-γ, ovalbumin IL-13, and itself, in the fifth *year-of-life*, the first of these with only intermediate posterior, in those who did not have *wheeze* in the fifth *year-of-life*. The corresponding DBN generated using all complete records finds no edges among any of the ovalbumin IL (not shown), but an interesting DBN was found by taking the log() of the interleukin values. The DBN in figure 5.31 indicates that ovalbumin IFN-γ, IL-10 and IL-13 led to *house-dust-mite IgE*.

The corresponding DBNs for tetanus, shown in figures 5.32 and 5.33, also display evidence of changes in interleukin regulation. There is no interleukin which appears to be of greater importance than any other. An association was published [169] between IL-γ and severe asthma in both mice and humans. We do not see it here, but this is

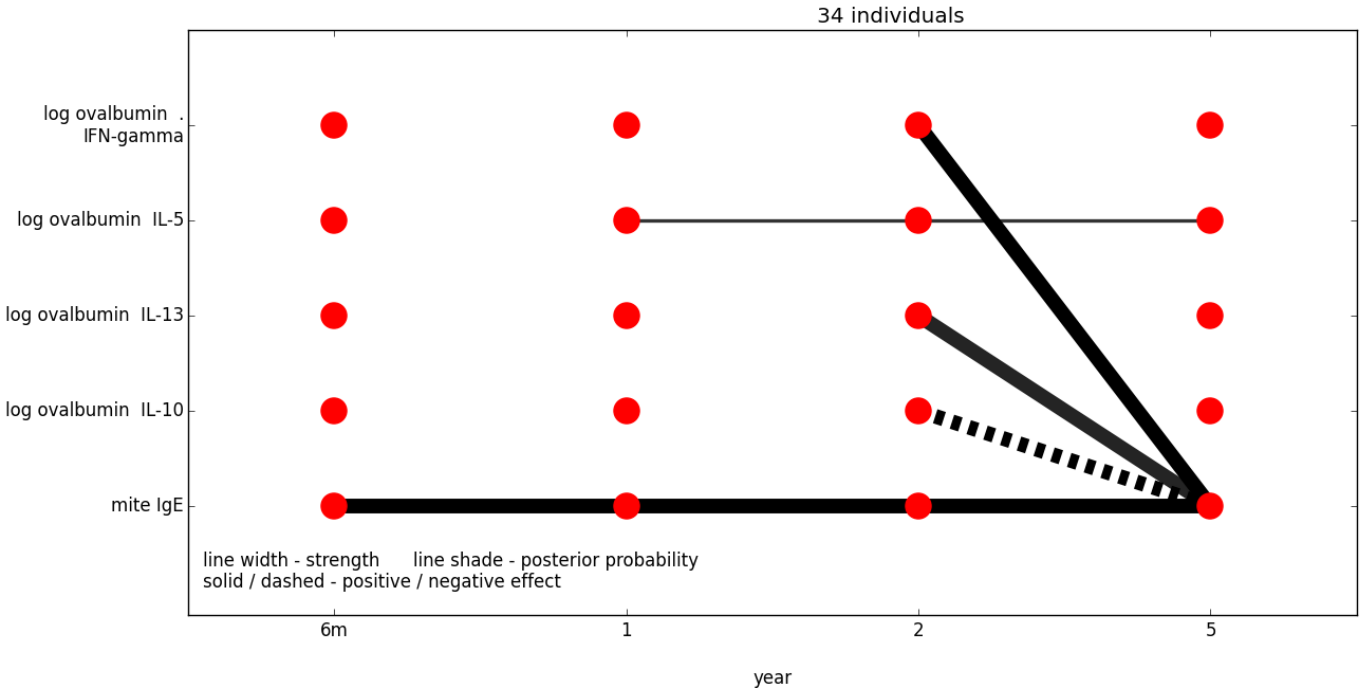


Figure 5.31: **(Log of) ovalbumin IL-13 and IFN-γ in the second *year-of-life* led to fifth-year *house-dust-mite IgE*, while (log of) ovalbumin IL-10 led against it. (Log of) ovalbumin IL-5 is ongoing from the first *year-of-life*:**

not surprising given its limited endotype (less than 10% of all asthma [3]) and the great deal of missing data among the interleukin variables.

The DBNs of interleukin were often empty and the few interesting networks we did find did not point to any clear culprit. What they did indicate is an apparent disregulation among interleukins, indicating a deeper underlying cause. IL-5 seemed to be involved in the differences between networks for both ovalbumin and tetanus. The significance of this is not known.

5.9 Outcomes

Much of this chapter was concerned with the interplay among infection, *wheeze* and *atopy*. Our networks do not support recent suggestions that *viral-LRI* leads to *airborne-atopy* or *atopic-wheeze*. In fact, there was a consistent pattern of infection being consequential rather than causal of *atopy* and *wheeze*. The possible exception is that *febrile-viral-LRI* from the second *year-of-life* onwards might increase the *atopy-number*

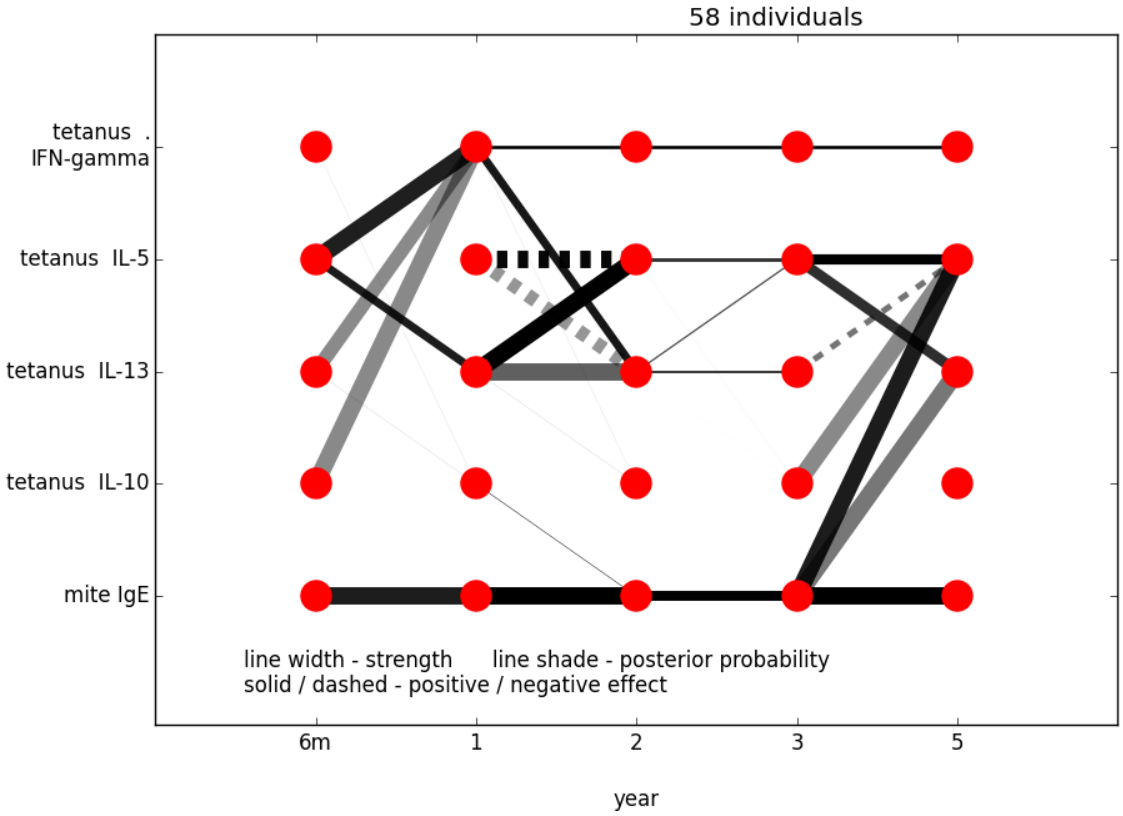


Figure 5.32: **Tetanus IL-10 in the first six *months-of-life* led to tetanus IFN-γ, while in the third *year-of-life* it led to tetanus IL-5 in the fifth *year-of-life*. All edges from tetanus IL-10 have intermediate posterior. *House-dust-mite IgE* in the third *year-of-life* led to tetanus IL-5 and to tetanus IL-13 in the fifth *year-of-life*, the latter with intermediate posterior. In common with figure 5.33, there was also interaction between tetanus IL-5, 13, which in the first six *months-of-life*, lead to tetanus IFN-γ:**

in those who are already *atopic*, but this might also be due to an *atopy*-inclined immune system being more likely to generate a fever.

As discussed in section 5.5.4, our analysis cannot address questions concerning re-modelling, specific *virii*, or infection in the early *weeks-of-life*. Within these caveäts, our model inferred the following :

1. *Wheeze* led to more *wheezy-LRI* (figures 5.1 to 5.10) and *viral-LRI* (figure 5.12).

2. *Airborne-atopic-wheeze* led to *wheezy-LRI* which would otherwise have been *URI* (figure 5.11).

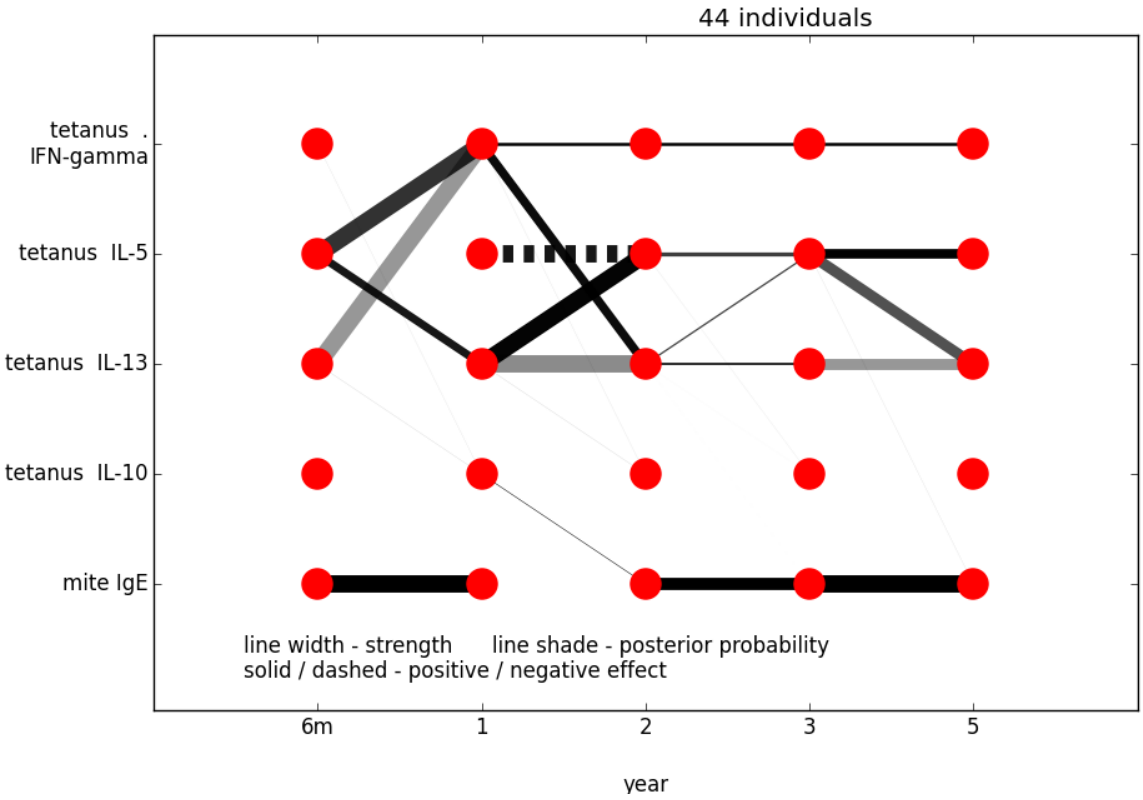


Figure 5.33: **Among those who did not exhibit *wheeze* in the fifth *year-of-life*, interaction between tetanus IL-5, 13, which in the first six *months-of-life*, led to tetanus IFN-γ:**

3. *Airborne-atopy* led to fewer *viral-URI* and more *severe-viral-LRI*. Of most importance were the *wheezy-viral-LRI* (figures 5.14 and 5.15).

4. *Nonatopic-wheeze* frequently resolves while *atopic-wheeze* does not unless the underlying *atopy* resolves (figures 5.5 and 5.6). We suspect that other forms of inflammation also prevent the resolution of *wheeze*, but that this dataset is too biased towards *atopy* to reach reliable conclusions about this.

5. Increasing *aeroatopy-number* in the first *year-of-life* led to *severe-viral-LRI*, and not the other way around. Biologically this means that a tendency for increasing numbers of *atopic* sensitivities also led to more *severe-viral-LRI*.

6. *Febrile-viral-LRI* did not cause *airborne-atopy* but, from the second *year-of-life*

onwards, led to higher *aeroatopy-number* if *airborne-atopy* was already present (figures 5.22 and 5.24). We cannot determine if this relationship is indicative or causal, so it is plausible that protecting against *febrile-viral-LRI* from the second *year-of-life* onwards might lead to a drop in *atopy-number*. This issue requires a specific intervention study.

7. We found no connection between *nonviral-LRI* and *wheeze*, *atopy* or *airborne-atopy* (generally not shown but see figure 5.22). We cannot rule out a lack of statistical power.

8. *IgE* dynamics are fundamentally different between those who later exhibit *wheeze* in the fifth *year-of-life* and those who do not. The former have much less cross-linking between different *IgE*s, apart from *house-dust-mite IgE* being dependent on earlier values of other *IgE*s. Conversely, the latter have *house-dust-mite IgE* driving the *IgE*s of other allergens, especially *cat*, *couch* and *rye*, although *peanut-IgE* in the first six *months-of-life* led strongly to *house-dust-mite IgE* in the first *year-of-life* (figures 5.25 to 5.28).

9. *Atopy-number* initially led to and was later led to by *house-dust-mite IgE*. The only other *IgE* to visibly interaction with *atopy-number* was *peanut*, which led to it in the first six *months-of-life* (figure 5.29).

10. Interleukin dynamics are fundamentally different between those who later exhibit *wheeze* in the fifth *year-of-life* and those who do not. The differences include interaction with *house-dust-mite IgE* (figures 5.31 to 5.33).

Chapter 6

Discussion

6.1 Identification of individual endotypes by classification

Asthma is a complex disorder, comprising multiple endotypes with differing causes, triggers and aetiologies. Contributing factors are generally predictive of only their own particular endotype, adding further to the difficulty of predicting *wheeze* in the fifth *year-of-life*. This is especially true when the endotype in question is such a small subset that the contribution to the general condition becomes swamped by noise.

The flipside is that we could use this selectivity to identify useful endotypes. It is a trivial matter to specify subsets of a general case, but only some are worthwhile. We were able to identify meaningful endotypes because their predictors did not just score AUCs predicting them that were higher than they did predicting the general case, but *predictably* higher, within a narrow margin for error. The clearest example of this was the known [20] second-year predictor, (log of) *house-dust-mite IgE*. Despite being the best predictor of fifth-year *wheeze* in the CAS study (*atopy-number* was constructed from CAS variables rather than being in CAS itself), its predictive ability was exclusive to the very specific *multi-allergen* endotype of *atopic-wheeze* requiring *atopic* sensitivity to at least one of *cat*, *peanut*, *couch* or *rye*.

6.2 Associations between the NP microbiome and asthmatic outcomes

Our study of the NP microbiome in the context of fifth-year *wheeze* and *atopy*, presented in chapter 4, considered two general cases. The first was samples taken from individuals during the course of a respiratory infection during the first 90 *days-of-life*. Dual qq-plots, *i.e.* one qq-plot for cases and one for controls on the same pair of axes,

associated higher levels of *Haemophilus* with *nonatopic-wheeze* in the fifth *year-of-life*, while *Staphylococcus* was associated with increased risk for *atopic-wheeze* against which *Moraxella* was found to be protective. The axes were the relative abundances of one of the *genera*, *Alloiococcus*, *Streptococcus*, *Staphylococcus*, *Haemophilus*, *Moraxella* and *Corynebacterium*, that typically constituted the NP microbiome.

The second general case was samples taken during the first seven *weeks-of-life* (< 50 days) when the individual did not have a respiratory infection. The dual qq-plots found only one signal, that for *Streptococcus* and fifth-year *wheeze*. This is not the first evidence for such an association [7], although the predictive signal from *Streptococcus* became noteworthy only when specific *atopic*-allergens were considered. Of particular interest were the allergens associated with the *multi-allergen* endotype presented in section 3.4. All but *rye* were very well predicted with AUCs in the 80% range. Our exclusivity index from chapter 3 did not find it to be exclusive to any given endotype, but given the small number of cases this may well be due to a lack of statistical power.

In both general cases the *genera* of interest generally indicated their signals regardless of which other *genus* they were plotted with, with very few exceptions. We were prepared for more complication than this, imagining interaction between *genera* with regards to pathological conditions. As it was, the evidence indicated that specific *genera* had certain associated risks.

6.3 Results from ARTIVA-derived DBNs

High-powered cohort studies like CAS which include diverse data types typically include a range of different, non-Gaussian probability distributions. Bayesian tools typically assume Gaussian distributions, with results often better than might be intuitively expected [123–128], although it has still been shown to be a significant limitation on performance in some instances [115, 117, 132].

The Gaussian prior makes this assumption about the model coefficients, rather than the data itself. For continuous data this is a gentler assumption, allowing the Gaussian process prior to render the calculation tractable [8], without inherent conflict with the

data. However for discrete and binary data the implicit assumption of continuous data is fundamentally problematic. We addressed the issue with the long-known procedure of Albert and Chib [9], as discussed in subsection 5.2.2. Future extensions might include implementing logistic regression for binary variables instead of defaulting to linear regression.

Additionally, transcending the Gaussian distribution assumption also allowed us to study the interactions between factors without taking subsets of our already limited sample size. We could, for example, consider infections with and without atopy, without having to consider the *atopic* and *non-atopic* samples separately, by multiplying the number of infections by zero or one, according to the atopic status of the participant in question. This sometimes led to very limited numbers of cases for some binary conditions, such as *aero atopic-wheeze* in the first *year-of-life*. We present the number of cases for important binary conditions in appendix C.

In addition to the edge weights (dependency strengths), *ARTIVA*'s output also includes the edges' posterior probabilities. They correspond to the posterior probability that the edge is present, but we have found it helpful to interpret it as the (approximate) fraction of samples to which the relationship applies. Doing so allowed us, in combination with biological understanding, to identify relevant subsets and conditions *e.g.* the intermediate posterior associated with atopy led us to consider airborne atopies separately. In this way we were led to useful conditions and subsets by the data, instead of needing to consider every possible combination. Recall our requirement that networks only contain edges of near-unity posterior. Instead of simply eliminating edges with posteriors below a given cutoff, we willing incurred the liability of finding networks whose edge posteriors were all sufficiently high. This led us to unexpected consequences of this model for the effect of *atopy* and *wheeze* on infections' ability to get into the lungs. Not only was this the most natural way to understand the negative effect on *URI*, but it also generated the most likely edges on the best graphs.

While other feasible approaches may well exist, a guided approach such as ours was certainly necessary. The highly heterogenous nature of CAS variables, combined with

the correlations expected among variables selected for a common association, and indeed sometimes with overlapping definitions, ruled out the effectiveness of straightforward "all-at-once" methods. However even our use of intermediate posteriors required a choice of starting network, and in this we were led by some of the biological questions currently discussed in the asthma field [6, 16, 20], focussing on the interplay between *atopy* and infection.

Both of these guides required biological insight on the part of the researcher. We make no apology for this, it being appropriate for reasons other than necessity for, despite our data-centric approach, this was at heart a biological problem.

Our networks concerning *atopy* and infection have consistently indicated, at least in this model, that it is the former which led to the latter, and not the other way around. Even the one possible exception, *febrile-viral-LRI* leading to higher *atopy-number*, required the copresence of *airborne-atopy* as indicated in figure 5.24. Whether this edge indicates a genuine effect or merely an indicative relationship, the infection was not the seminal factor. Enquiries into the development of *IgE*-titres revealed highly disparate networks between those who went on to develop fifth-year *wheeze* and those who did not. Nonetheless, *house-dust-mite IgE* was shown to have played a pivotal role, being the child nodes in the non-*wheeze* networks and the parent nodes in *wheezy* ones. Figure 5.29 further indicates that (log of) *house-dust-mite IgE* was stimulated by *atopy-number* but also led back to it.

Important findings, especially those regarding infection, from this *ARTIVA*-based model were supported by the χ-squared test, but it is nonetheless prudent to indicate the limitations of this model. Its inherent linearity and year-by-year time-resolution leave it prone to missing relationships that are highly non-linear or whose time-scale is much shorter than one year. This latter point is especially important for data in early infancy, where significant variables were found within the first few *weeks*- and *months-of-life*. Addressing the former point might open the door to a new form of overfitting, but that might be overcome with an adaption of Albert and Chib's augmentation [9] or with information theory-based methods [174–177].

Chapter 7

Conclusion

The complexity of asthma pathogenesis is undeniable. A large part of this is the substantial number of endotypes whose variations cover predictors, severity, age-of-onset, *atopy*-status, medication response, and triggers. Nor are all triggers *atopic*, with cold and exercise triggering attacks in some cases [15].

We have demonstrated, through our novel exclusivity index, that some biologically important variables were predictive of specific endotypes only, while other important variables, notably *severe-* and *febrile- LRI*, were not endotype specific but predictive of *wheeze* in general. This raises questions concerning the causal status of infection. If, as some researchers believe [178, 179], *viral* infections play a causal role then either the outcome is very dependent on the specific virus or the pathogenic effects are dependent on other variables, such as the theorised interaction between *viral-LRI* and *atopy* [16].

However, our finding in subsection 5.5.1 was that *viral-LRI* were not causal of *wheeze* or *atopy*, but caused by them, which sits well with the non-exclusive nature of infection as a predictor. The only active role indicated by our model was for *febrile-viral-LRI*, as a preserver of *airborne-atopy* from the second *year-of-life* onwards. The simplest interpretation is that *atopy*-prone immune systems are inclined toward fever, but the lack of a negative signal from *mild-viral-LRI* is also suggestive. If it is not mere association, the network in figure 5.4 indicating that *wheeze* with *airborne-atopy* does not resolve then suggests that there might be a long-term benefit from preventing *febrile-viral-LRI* among *atopic* children.

We have found that asthma needs to be understood in terms of its individual endotypes which may be analytically determined *via* the exclusivity index (equation 3.5), and that one of the best predictors of asthma, (log of) *house-dust-mite IgE* in the second

year-of-life [19], is exclusive to a particular endotype which is predictable from non-*ARI* NP microbiome samples from the first seven *weeks-of-life*. Other associations from *ARI* microbiome samples lay between various *wheeze*-related outcomes and certain *genera*. These were detected by dual qq-plots but were not useful for prediction. In the case of *Staphylococcus* the association was argued to be indicative only.

We also found that the derived variable *atopy-number*, the number of allergens to which an individual was *atopic*, was a strong and exclusive predictor from the first two *years-of-life* of fifth-year *aeroatopic-wheeze*. Interestingly, counting the total number of *atopic* triggers was more effective than just counting the airborne ones.

We have also shown that Gaussian prior-based models like *ARTIVA* can be extended for use with non-continuous data. Our augmented model then found that the occurence of (*viral-*) *LRI*s in early childhood and especially in the first *year-of-life*, was led to by *airborne-atopy* and *wheeze* instead of leading to them. Indeed, the effect was sufficiently pronounced to lower (*viral-*) *URI* by a corresponding amount. The conclusion is not so strong from the second *year-of-life* for the reasons discussed above, and the edges from *febrile-viral-LRI* might be due to a merely indicative relationship. *Non-aeroatopic-wheeze* was of no detectable relevance to *wheeze* or asthma.

Appendix A

Multiple predictors and overfitting

A.1 Multiple predictors and overfitting

Overfitting is an ubiquitous factor limiting machine learning performance. It is characterised by the counter-intuitive phenomenon of a classifier's performance degrading in response to the inclusion of additional predictors, and caused by the machine learning algorithm learning the training set to such excessive accuracy that it starts to fit the noise in the signal. An alternative characterization is that testing on the training set gives a better result than testing on an independent testing set, which is the reason such classifiers are typically tested using *cross-validation* or *bootstrap* methods. Overfitting is also a disincentive to add nonlinearity or additional features to a model without specific motivation.

To study overfitting we generated multiple simulated predictors. Each simulated predictor was generated by two Gaussian distributions, corresponding to each of the classes, positive and negative, of the binary classification. Each of these class-dependent distributions had its own mean and variance, where the means would normally differ between the two classes.

When more than one simulated predictor was to be considered at once their class-dependent distributions were generated with pre-determined correlations (0.5 unless stated otherwise) with the corresponding distributions of the first simulated predictor. We could, in principle, have specified all correlations among all predictors but saw little benefit for the extra effort it would cost the user to specify the correlation matrix. We found that just specifying a correlation with the first predictor was sufficient to explore the effects of correlation. We constructed the required correlation matrices with R's

gen_ corr( ) function.

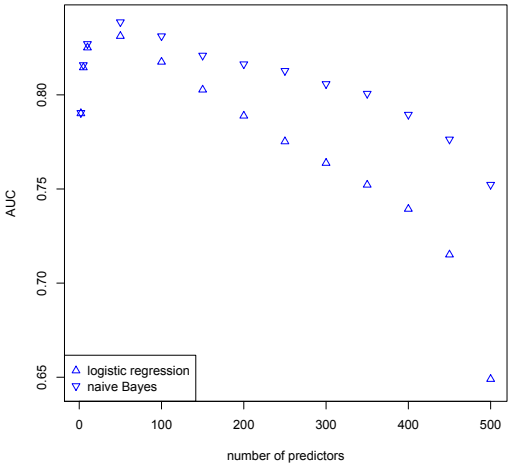
Our simulations, illustrated in figure A.1 illustrate the effects of overfitting on both logistic regression and naïve Bayes classifiers. As expected, logistic regression performance decayed to that of random guessing with large numbers of predictors, while naïve Bayes suffered some degradation but retained asymptotic predictive power (see subfigures A.1a, A.1b). This is consistent with naïve Bayes being theoretically immune to overfitting [180], subject to the quality of its implementation [113, 115, 116], while there is no corresponding theorem for logistic regression. Neither classifier demonstrated any degradation in performance due to over-fitting when the simulated predictors were generated with no correlations among them, as shown in figure A.1c.

Figure A.2 demonstrates that naïve Bayes outperformed logistic regression for multiple CAS predictors. While naïve Bayes performance plateaued to a weighted mean of individual AUCs, that of logistic regression decayed to .50 with increasing numbers of predictors. (We chose the variables with the most Gaussian distributions, as measured by the Shapiro-Wilks test, and added them to the classifier in both ascending and descending order. We attribute the relative performance between ascending and descending Shapiro-Wilks index to the distributions of the best predictors.) While prudence should always be shown when considering different data sets, the highly heterogeneous nature of the included predictors gives some confidence that this generic result should be widely applicable.

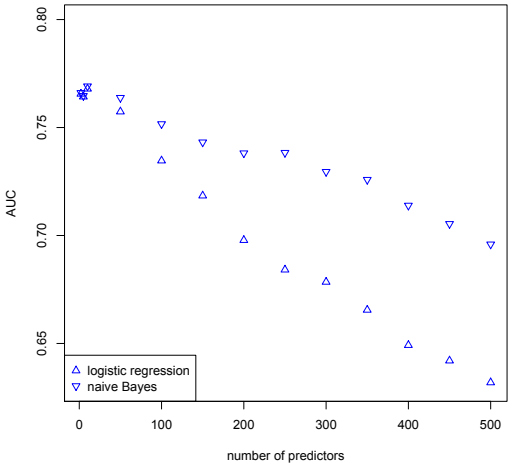
A.2 Combining predictors of different quality

We then studied the effect of combining predictors of different quality. We controlled this quality by adjusting the difference between the means of the class-dependent distributions, and also by adjusting their standard deviations. A greater difference in their means unsurprisingly gave a higher AUC while larger standard deviations led to a lower one.

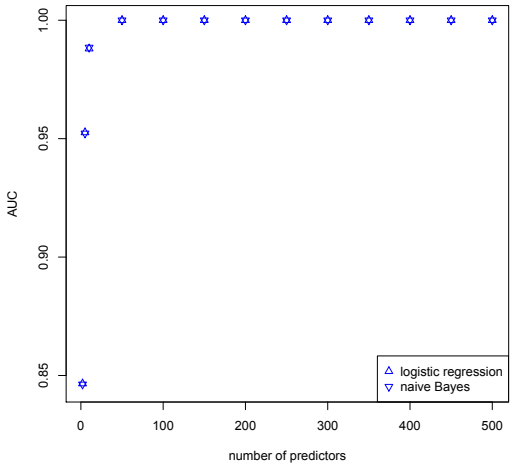
The top two graphs in figure A.3 show the AUCs found by both logistic regression and naïve Bayes for ensembles of two and three simulated predictors. In the top-left



(a) **Overfitting in the presence of correlations was more detrimental to logistic regression than to naïve Bayes:** The correlation between the class-dependent distributions of the first simulated predictor and the corresponding distributions of the later predictors was 0.5.



(b) **Stronger correlations among multiple predictors were detrimental to both classifiers:** The correlation between the class-dependent distributions of the first simulated predictor and the corresponding distributions of the later predictors was 0.9.



(c) **Overfitting did not affect simple classification in the absence of correlations:** These simulated predictors were uncorrelated.

Figure A.1: **Overfitting as demonstrated by simulated predictors:** All simulated predictors had a difference in mean of one and a variance of one for both class-dependent distributions. The plotted points are spaced at intervals of 50, with additional plots for two, five and ten predictors.

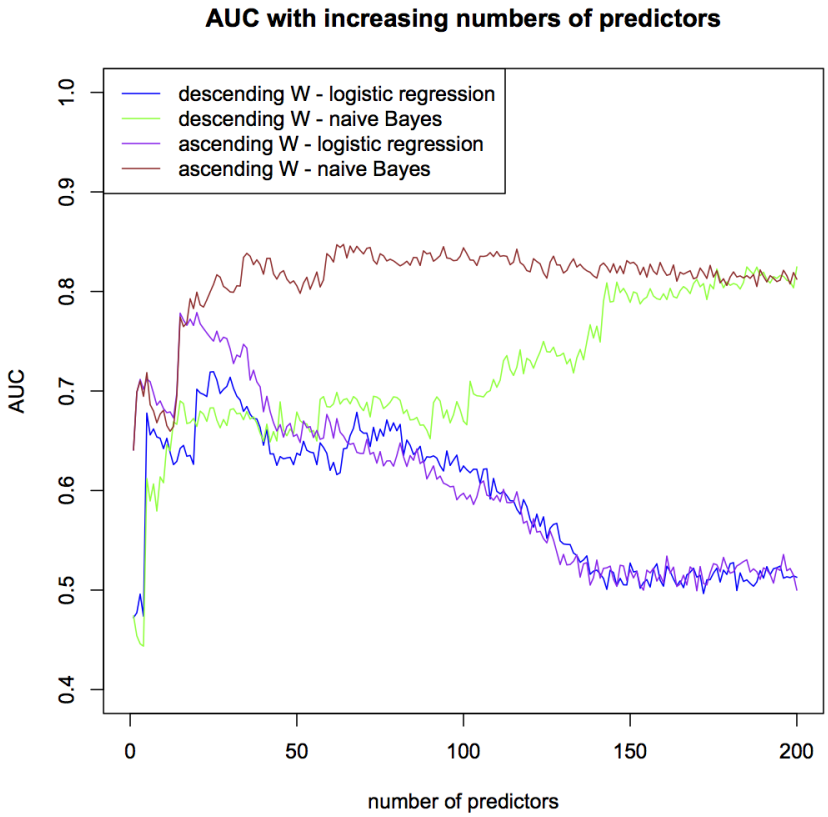


Figure A.2: **Classifier performances with increasing numbers of CAS predictors:** The included predictors are those with one of the top 200 Shapiro-Wilks W scores. The four lines indicate the AUC as a function of the number of predictors. Logistic regression and naïve Bayes classifier performances are plotted for each of increasing and decreasing AUC. The logistic regression classifier has its peak performance at about 20-25 predictors and declines from then onwards with perhaps one hiatus, while the naïve Bayes-curves decline only marginally after reaching their peaks.

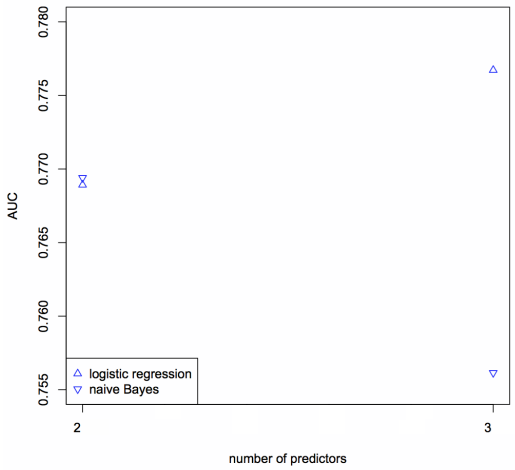
graph the third predictor had a lower AUC than the first two, effected by giving the class-dependent distributions of the third predictor a standard deviation which was double that of the first two. The logistic regression classifier performs better with the introduction of the weaker predictor while the naïve Bayes classifier is weakened. In the top-right graph the first predictor is the weak one, while the other two are stronger (standard deviations 2, 1, 1, respectively). Logistic regression performs better with the unmatched pair, but they both improve by the same amount when another strong predictor is added.

This general pattern is repeated by the bottom two figures, with logistic regression and naïve Bayes again performing equally well with a pair of equally good predictors in the left-hand graph in which both had a relatively low AUC, while logistic regression and naïve Bayes improved and declined, respectively but both classifiers were equally affected by the introduction of a third predictor with a higher AUC. In this case the difference in AUC was effected by the first two predictors each having a difference of three between the means of their class-dependent distributions, while the remaining predictor had a difference of one, while all class-dependent distributions had a standard deviation of three. We again found that adding an inferior predictor degraded the predictive performance of naïve Bayes, both in comparison to having fewer predictors and in relation to logistic regression.

It seems advisable to avoid adding poorer predictors to predictor ensembles when using naïve Bayes classifiers, but that this is not so important for logistic regression. However, the performance of logistic regression will degrade with large numbers of predictors, regardless.

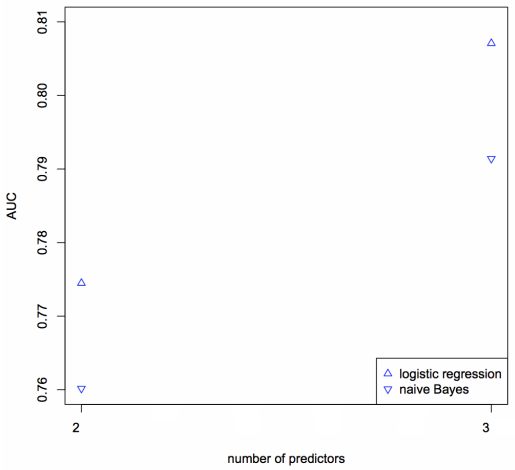
After completing this work we found a paper by Subramanian and Simon [162] which also used simulations to study over-fitting. They measured the difference in accuracy between testing on the training set and testing on an independent test set to measure overfitting in a range of classifiers, including logistic regression. Their approach found that more accurate predictors were less vulnerable to overfitting, and *vice versa*. Indeed, overfitting was observed among null predictors (AUC = .50), and also increased with smaller samples.

The degree of correlation and the quality of the individual predictors determined both the asymptotic performance of naïve Bayes and the rate of decline of logistic regression. We therefore advocate the practice [117] of choosing predictors according to a greedy algorithm, starting with the best predictors and proceeding in decreasing order of individual predictive power.



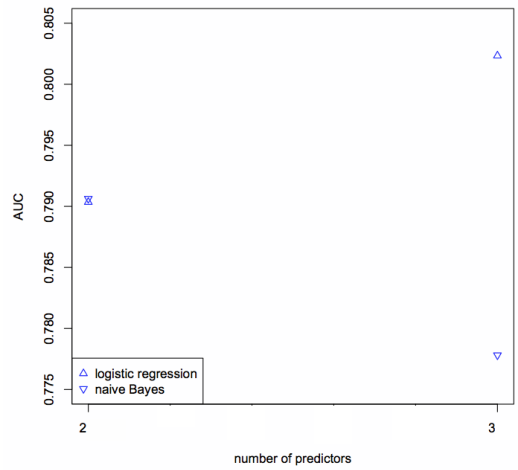
(a)

standard deviations	LR	NB
1, 1	.769	.769
1, 1, 2	.777	.756



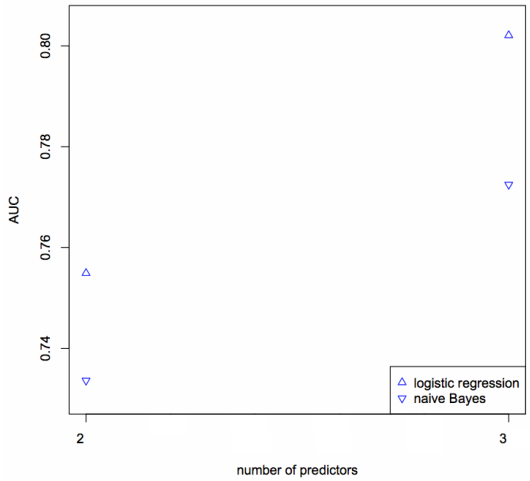
(b)

standard deviations	LR	NB
2, 1	.774	.760
2, 1, 1	.807	.791



(c)

standard deviations	LR	NB
3, 3	.627	.627
3, 3, 1	.773	.739



(d)

standard deviations	LR	NB
1, 3	.745	.726
1, 3, 3	.745	.711

Figure A.3: **Comparisons of simulations with two or three simulated predictors for both logistic regression and naïve Bayes:** The tables give the logistic regression (LR) and naïve Bayes (NB) AUCs for the standard deviations corresponding to the simulated predictors. For every predictor the difference between the means of its distributions is one. Comparing these graphs, it appears that the inclusion of a weaker predictor adversely affects the performance of the naïve Bayes classifier more than the logistic regression classifier.

Appendix B

Augmentation of integer data for Gaussian priors

The augmentation of Albert and Chibb [9] replaces every discrete dependent variable Y_i with a continuous variable Z_i, where the subscript i indicates the sample (participant). The value of Z_i is randomly generated from a truncated Gaussian distribution at every iteration of the Gaussian prior algorithm, with the boundaries of the truncation depending on the value of Y_i.

The simplest application is when the Y_i are binary. There is a single cutoff, y_0 say, to mark the difference between the two available values (0 and 1 in our work). The value of y_0 can be set arbitrarily, and remains constant throughout. It is common to set $y_0 \equiv 0$ but since we also deal with count data we found it more convenient to take $y_0 \equiv \frac{1}{2}$.

For a system of linear equations

$$\mathbf{A}\vec{X} + \vec{B} = \vec{Y}, \tag{B.1}$$

we make the replacement

$$Y_i \to Z_i \sim \mathcal{N}(\mu_i, 1), \quad \mu_i = \sum_j (A_{ij} X^j + B_i), \tag{B.2}$$

truncated according to

$$\begin{aligned} (-\infty, \frac{1}{2}) \quad & \text{when } Y_i = 0 \\ (\frac{1}{2}, \infty) \quad & \text{when } Y_i = 1 \end{aligned} \tag{B.3}$$

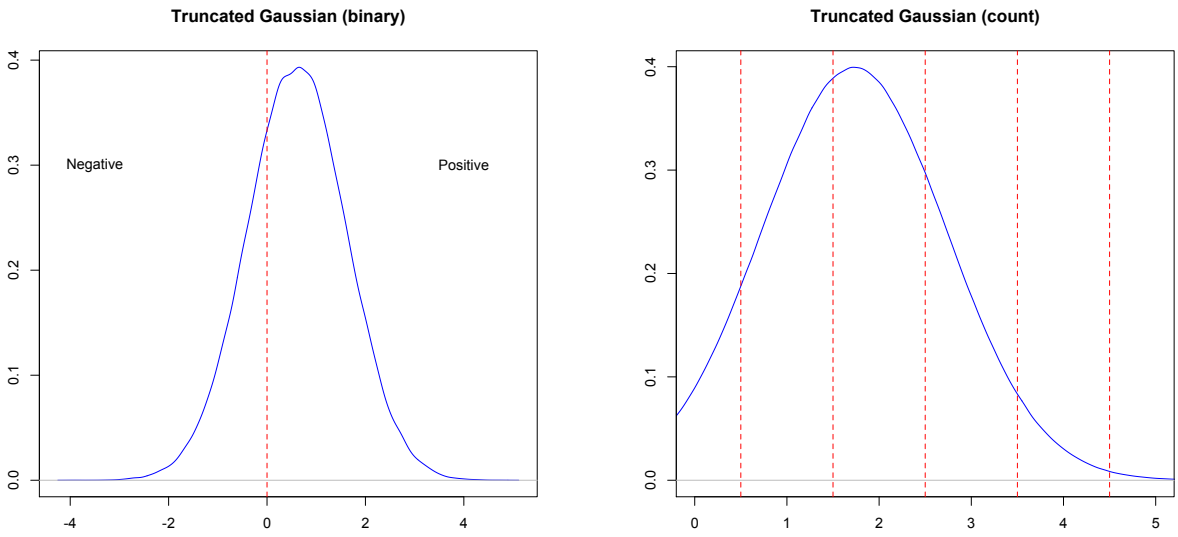


Figure B.1: The peak of the Gaussian in the left-hand figure is in the positive range. If the corresponding $Y_i = 1$ then the selected Z_i will be typically chosen around the peak of the distribution, or up against the left-hand-side of the cutoff otherwise. Similarly in the right-hand figure, the peak of the Gaussian is between one and two, with the Z_i distributed around this peak if the corresponding $Y_i = 2$, or up against the adjacent cutoffs, otherwise. This method allows the MCMC to also sample $\mathbf{Z}$ distributions.

This is illustrated in figure B.1.

When Y_i can be one of a larger set of discrete values it is necessary to have more than one cutoff. The exposition given in [9] treats the general case of Y_i corresponding to categories, and includes a step adjusting the cutoffs at each iteration. Since our discrete data are actually count data, where the unit interval is mathematically meaningful, we do not need this degree of generality and may set the cutoffs *a priori* at the half integers, *i.e.*

$$-\infty, \frac{1}{2}, \cdots, \mathbf{Y_i} - \frac{\mathbf{1}}{\mathbf{2}}, \mathbf{Y_i} + \frac{\mathbf{1}}{\mathbf{2}}, \cdots, \max(Y_i) - \frac{1}{2}, \infty \tag{B.4}$$

This is illustrated in figure B.1

Appendix C

Numbers of cases and controls for important conditions

Several of our analyses involve specification of precise conditions, so the numbers of cases and controls are provided here for the reader's reference.

condition	case/control/NA	year 1	year 2	year 3	year 4	year 5
atopy (ae)	cases	68	84	83	80	89
	controls	136	118	117	106	80
	NAs	2	4	6	20	37
airborne-atopy	cases	16	54	56	68	78
	controls	188	148	144	114	91
	NAs	2	4	6	20	37
wheeze	cases	67	63	59	54	56
	controls	139	143	147	145	141
	NAs	0	0	0	7	9

Table C.1: **Numbers of cases and controls for binary variables included in CAS and included in analyses**

condition	case/control/NA	year 1	year 2	year 3	year 4	year 5
airborne-atopy with *wheeze*	cases	8	22	22	27	30
	controls	196	180	178	159	139
	NAs	2	4	6	20	37
airborne-atopy without *wheeze*	cases	8	32	34	41	48
	controls	196	170	166	145	121
	NAs	2	4	6	20	37
wheeze without *airborne-atopy*	cases	57	40	37	25	20
	controls	147	162	163	161	149
	NAs	2	4	6	20	37

Table C.2: **Numbers of cases and controls for compound binary variables constructed from those in CAS and included in analyses**

Appendix D

Numbers of infections

Some of the networks in chapter 5 are based on the numbers of varioius kind of infection.

We have covered only the most important infections as a compromise to brevity.

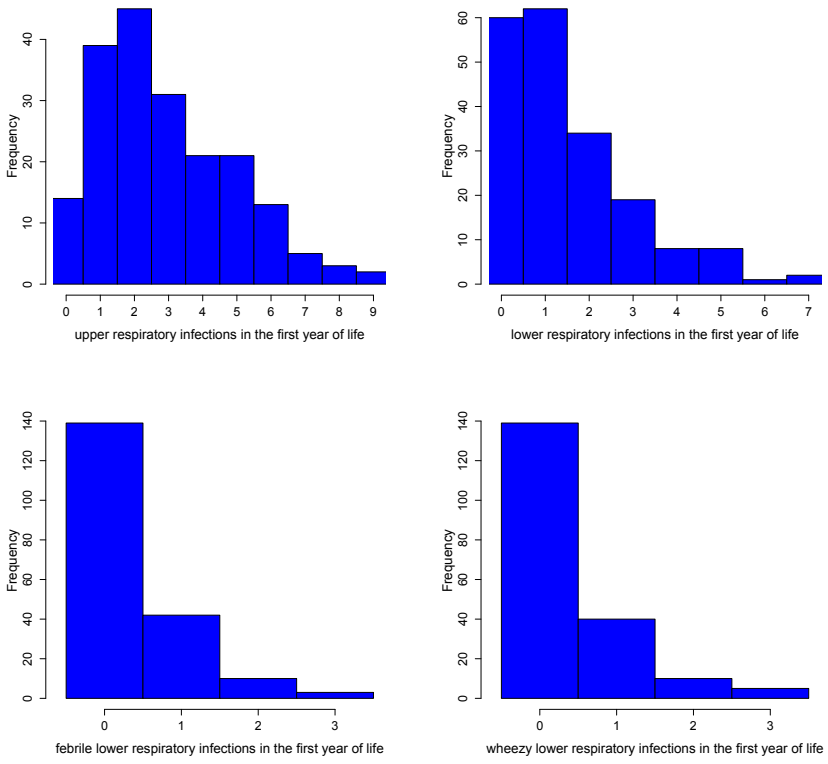


Figure D.1: **The number of important types of infection during the first *year-of-life*.** Infection numbers not shown can be inferred

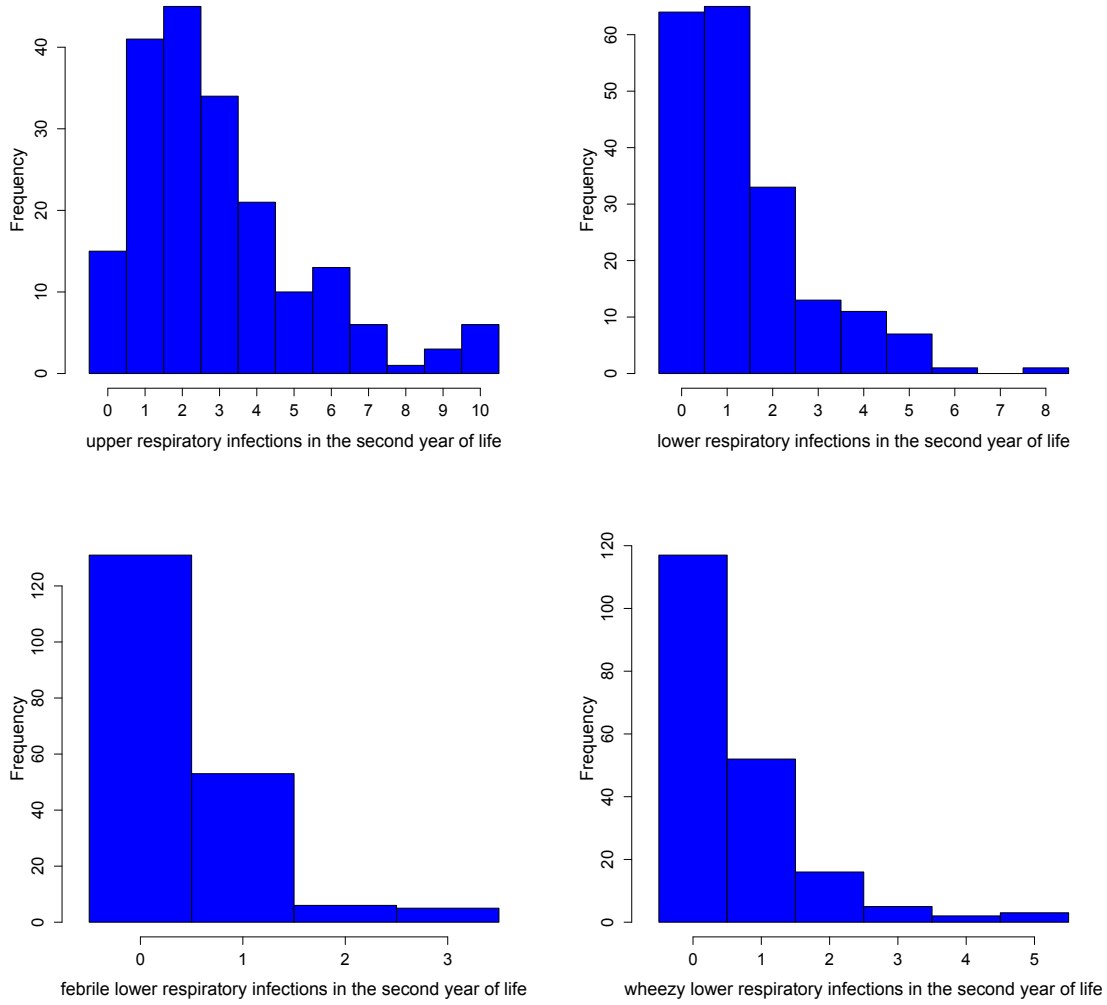


Figure D.2: **The number of important types of infection during the second *year-of-life*.**

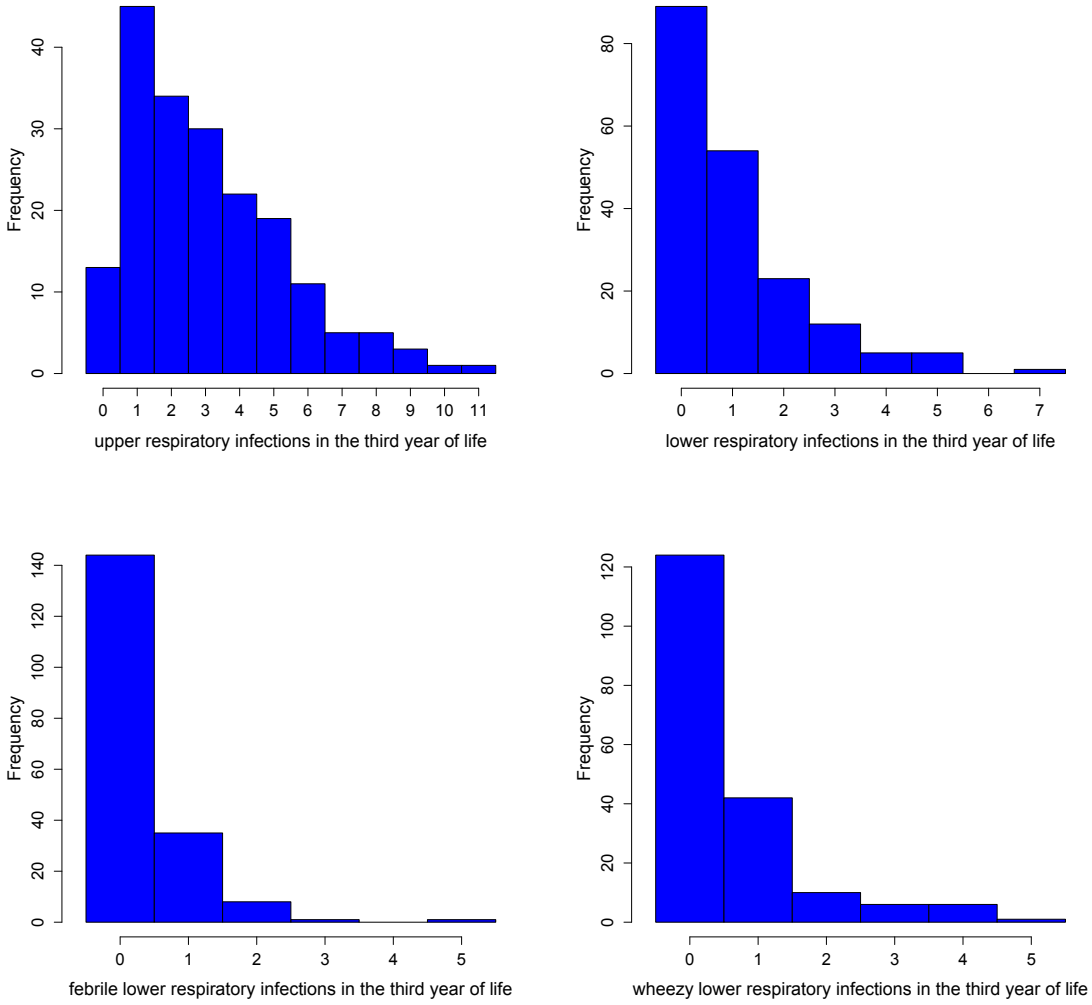


Figure D.3: **The number of important types of infection during the third *year-of-life*.**

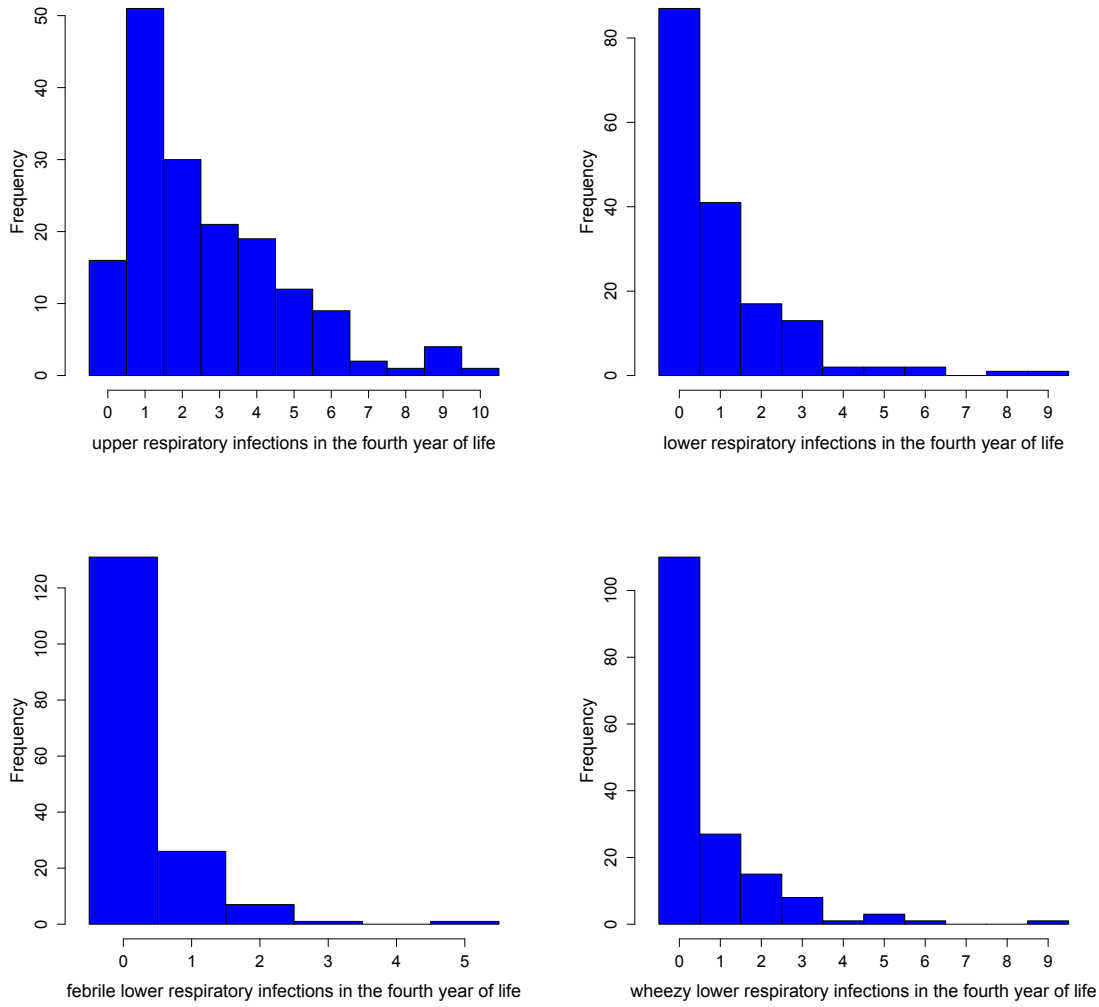


Figure D.4: **The number of important types of infection during the fourth *year-of-life*.**

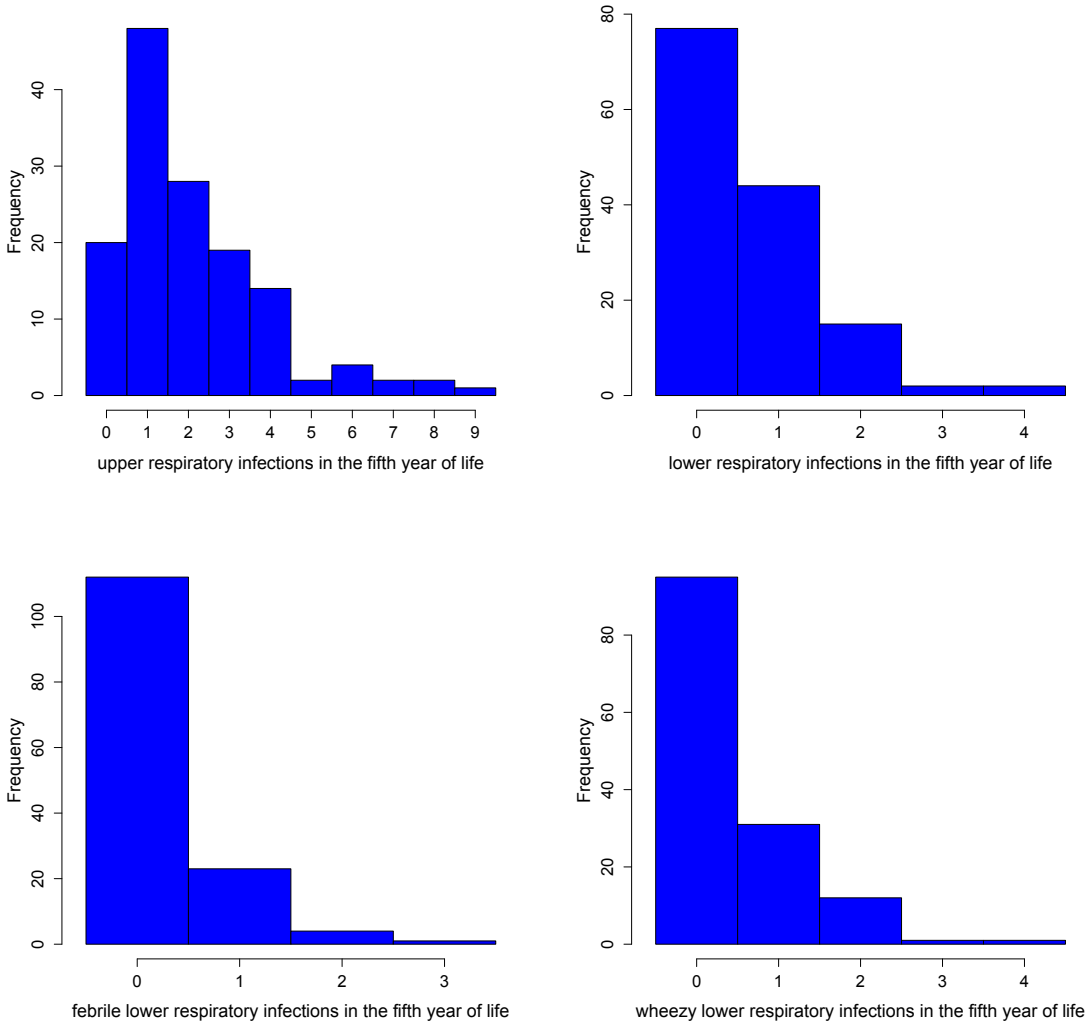


Figure D.5: **The number of important types of infection during the fifth *year-of-life*.**

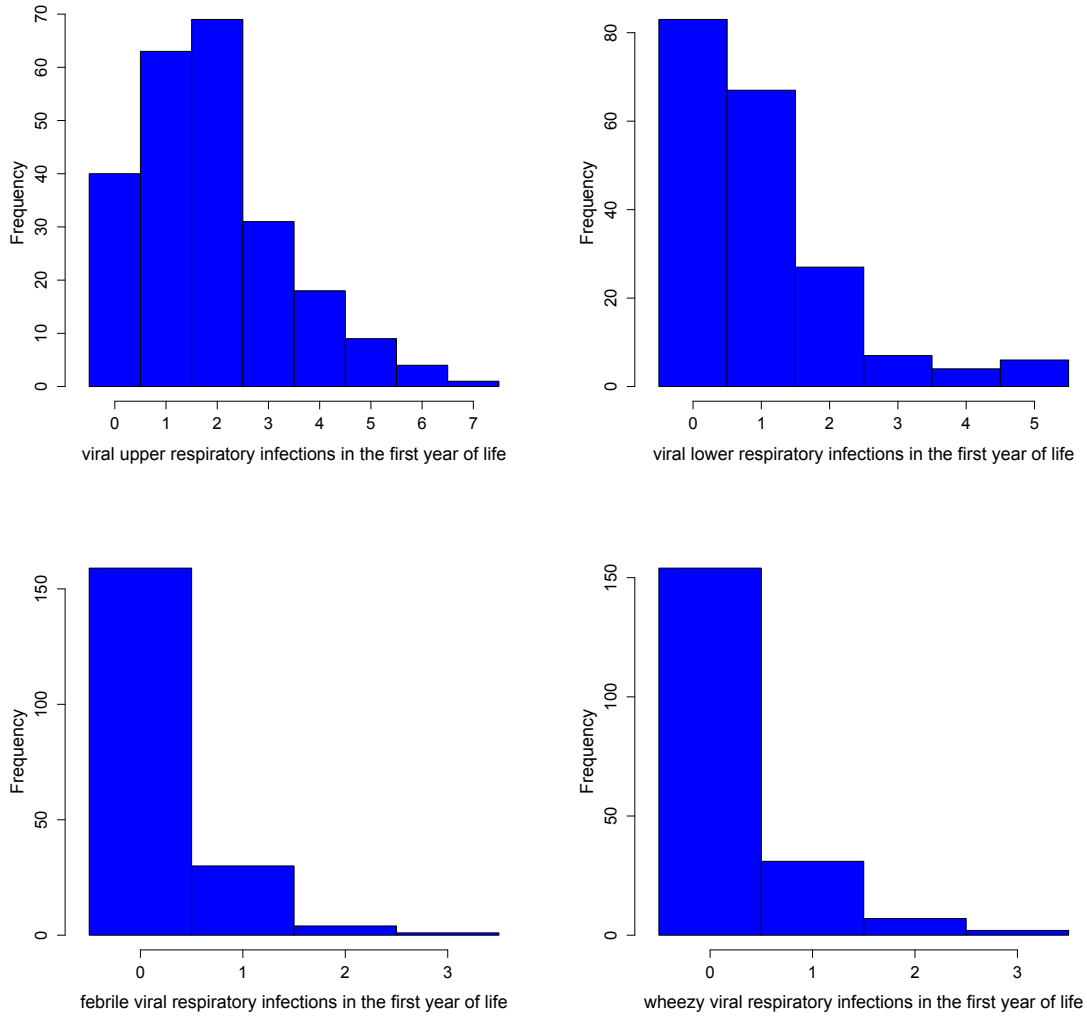


Figure D.6: **The number of important types of *viral* infection during the first *year-of-life*.**

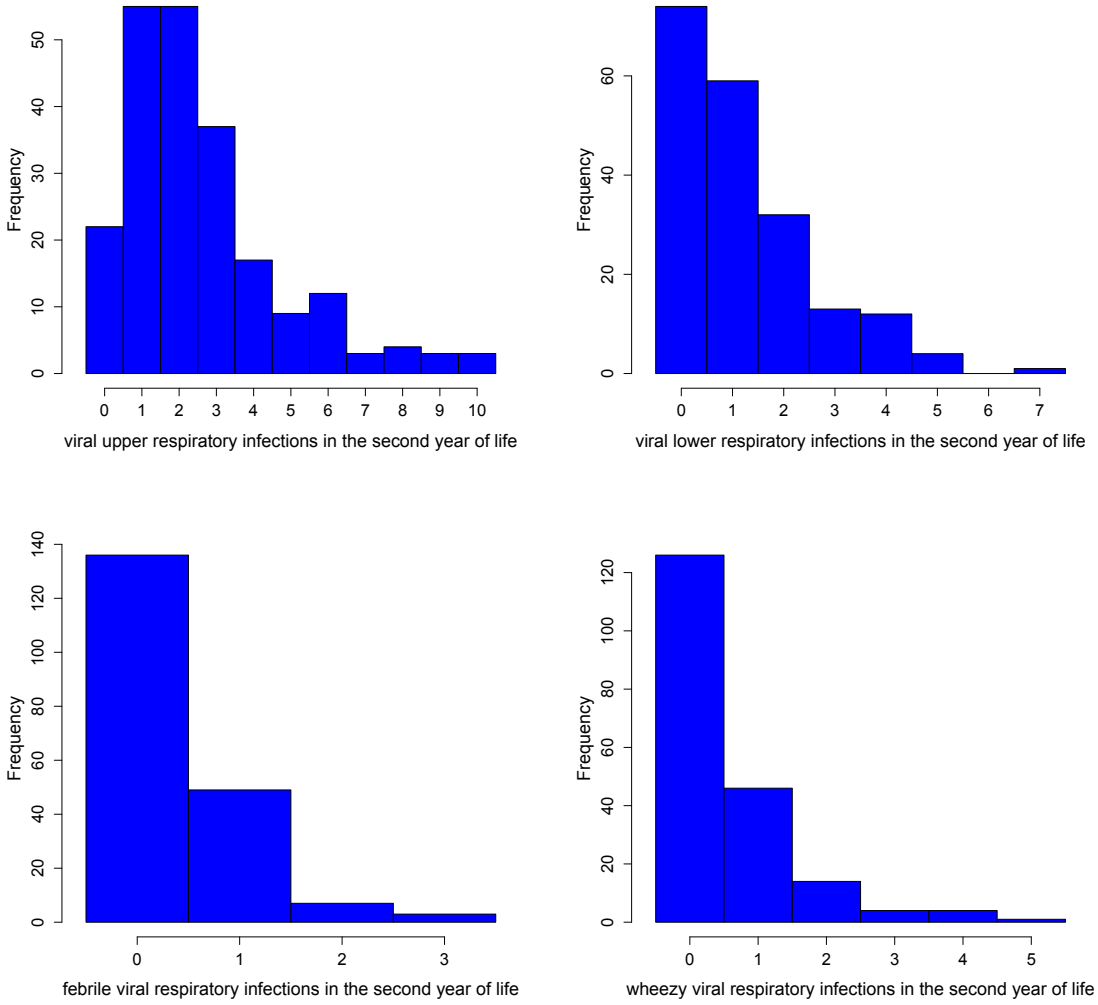


Figure D.7: **The number of important types of *viral* infection during the second *year-of-life*.**

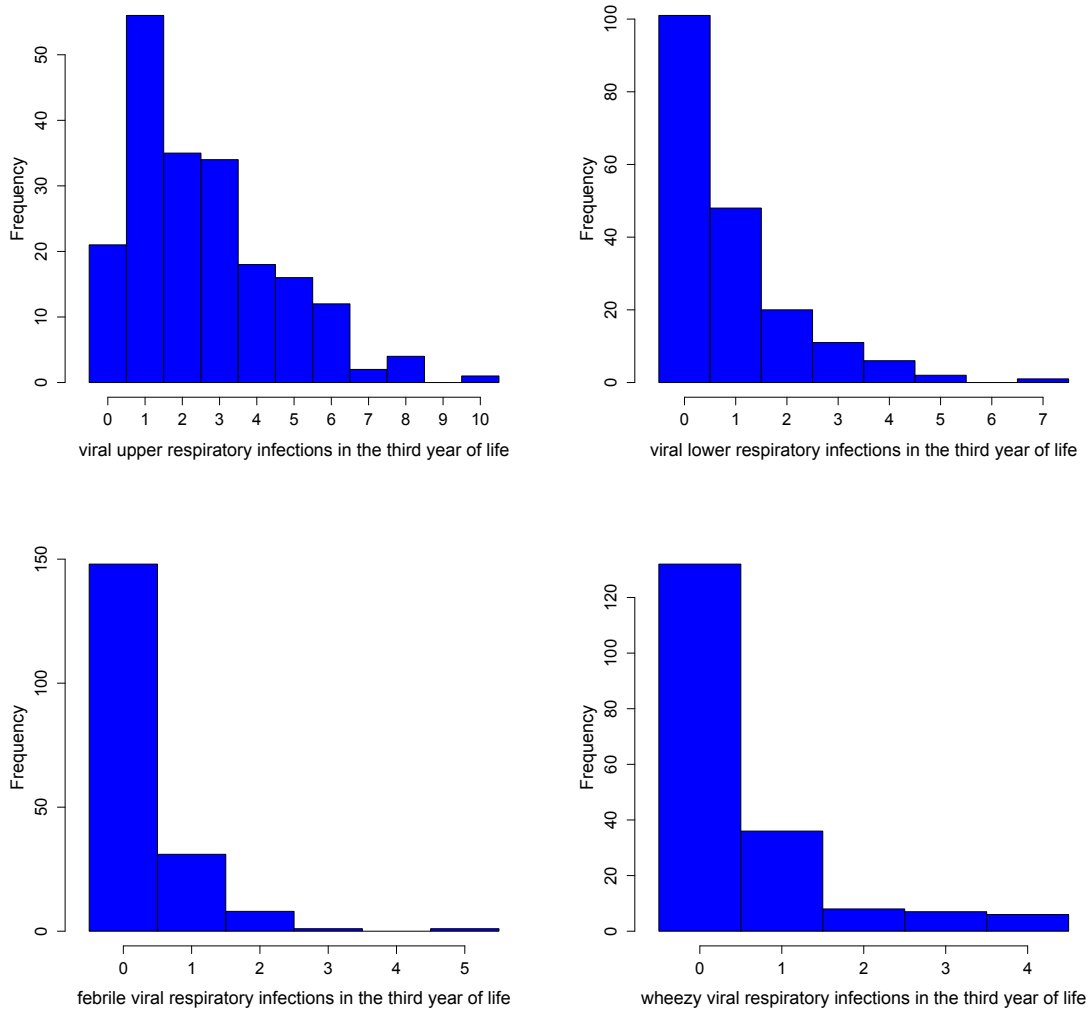


Figure D.8: **The number of important types of *viral* infection during the third *year-of-life*.**

Appendix E

QQ-plots to supplement discussion in section 4.2

E.1 QQ-plots illustrating the relationship between *Haemophilus* and fifth-year *atopic-wheeze*

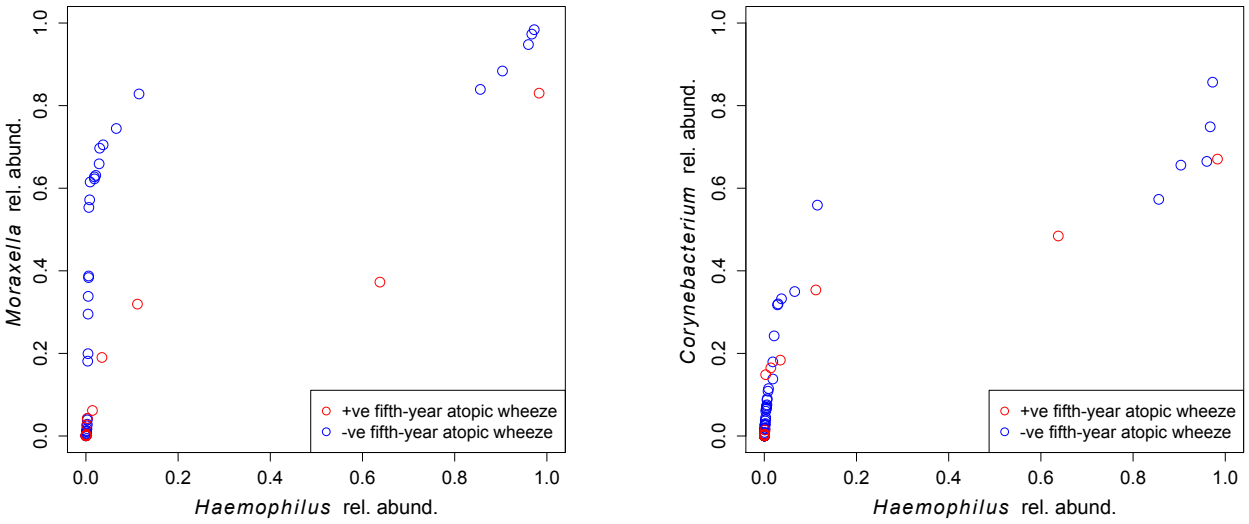
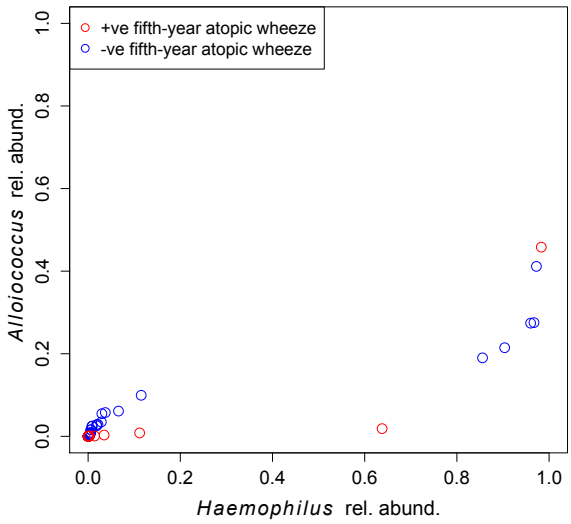
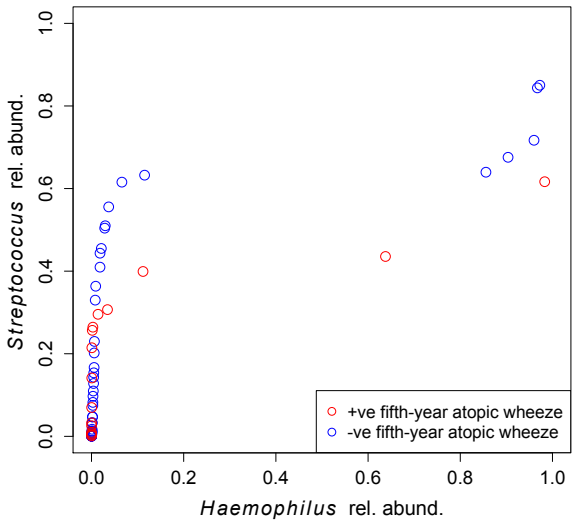
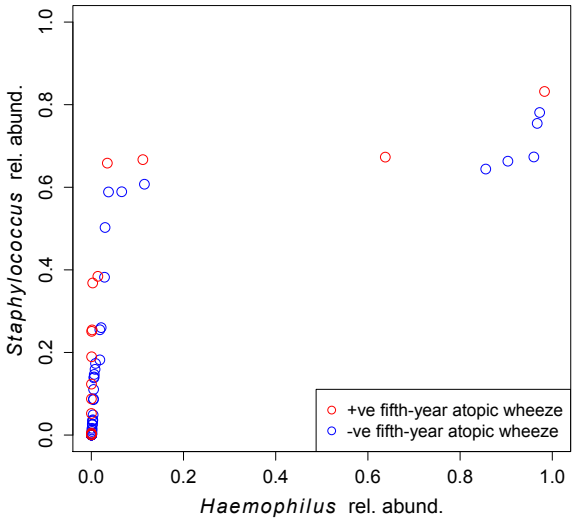


Figure E.1: **Remaining qq-plots demonstrating a link between an elevated relative abundance of *Haemophilus* in biomes from samples taken when the participant was suffering a respiratory infection, and *atopic-wheeze* in the fifth *year-of-life*:**







QQ-plots demonstrating a link between an elevated relative abundance of *Haemophilus* in biomes from samples taken when the participant was suffering a respiratory infection, and *atopic-wheeze* in the fifth *year-of-life*:

E.2 QQ-plots illustrating the relationship between *Moraxella* and fifth-year *wheeze*

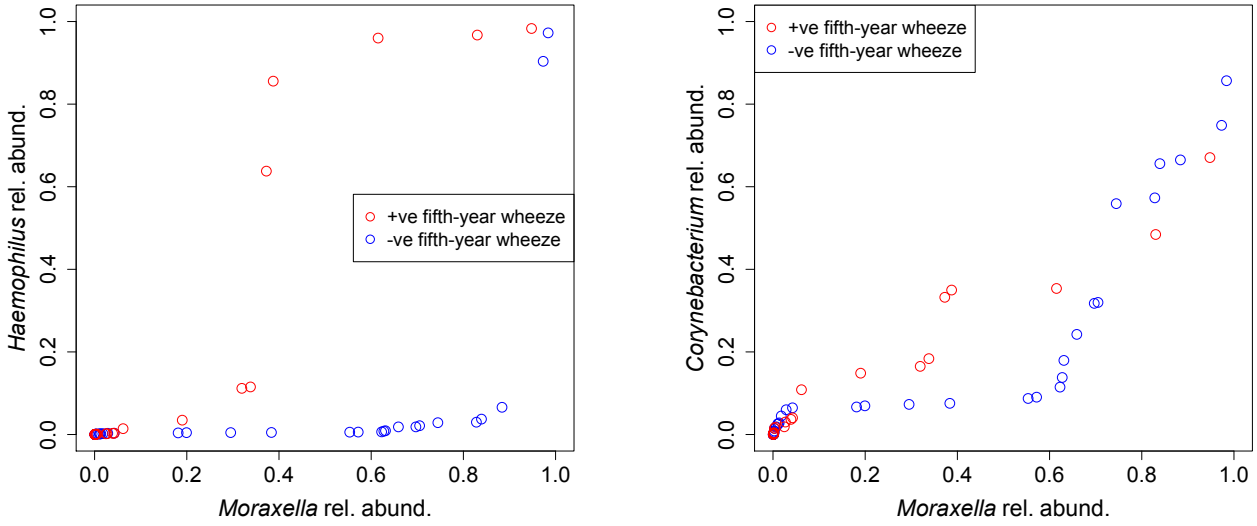
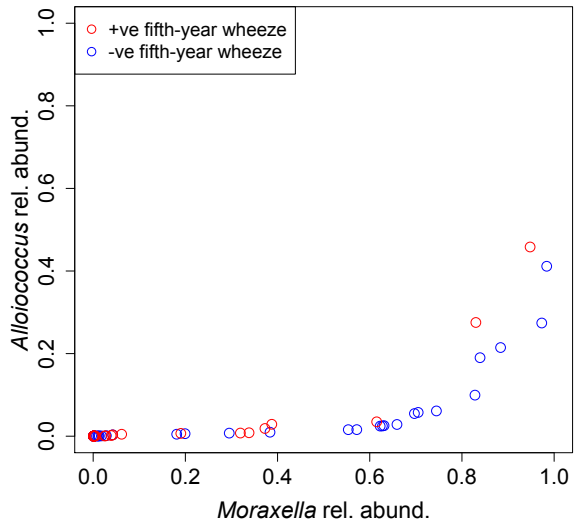
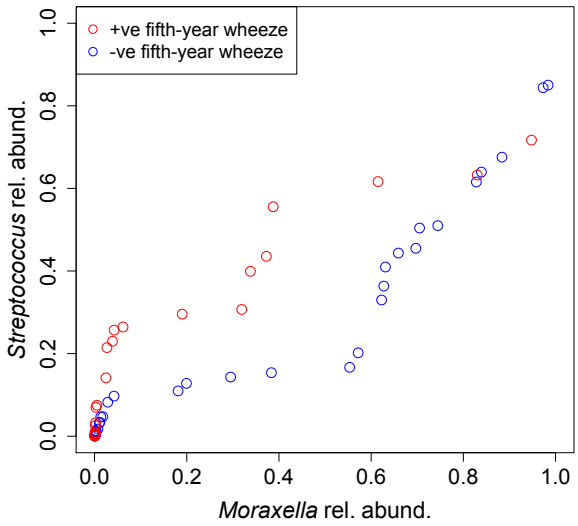
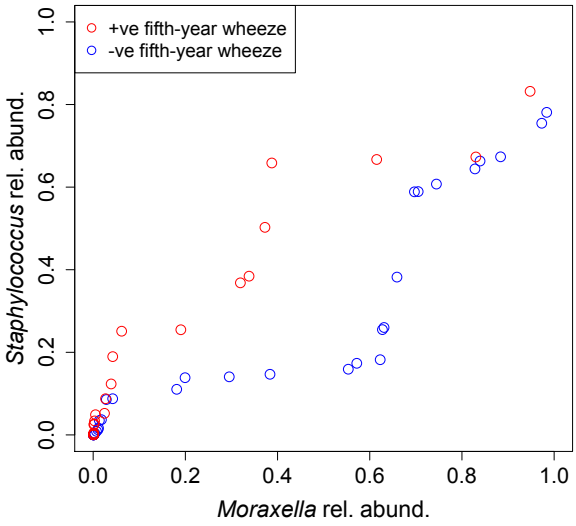


Figure E.2: **Remaining qq-plots demonstrating a link between an elevated relative abundance of *Moraxella* in biomes from samples taken when the participant was suffering a respiratory infection, and lack of *wheeze* in the fifth *year-of-life*:**







QQ-plots demonstrating a link between an elevated relative abundance of *Moraxella* in biomes from samples taken when the participant was suffering a respiratory infection, and lack of *wheeze* in the fifth *year-of-life*: The effect is less pronounced than for *atopic-wheeze* specifically, indicating that this effect is *atopy* related.

Appendix F

Method for acquiring *genera* abundances from the NP microbiome

The following is extracted with minimal modification from supplementary S1 of Teo *et al.* [7], which includes further details less relevant to this thesis.

F.1 Aspirate sample taking

Healthy NP Aspirates (NPAs) were collected from participants by study clinicians during planned visits at approximately 2 months, 6 months and 12 months of age, after the child had been free from any symptoms of respiratory illness for a period of at least 4 weeks. Parents were also asked to report to the study clinicians whenever the child showed symptoms of an Acute Respiratory Illness (ARI), at which point the family was visited within 48 hours by a study nurse who recorded clinical details of the infection and collected an NPA from the child. Clinical data recorded included the presence of fever, wheeze or rattly chest and any medications taken (including antibiotics). Each ARI was classified as either an LRI or a URI. The material in each NPA was divided into four aliquots and stored at $-80°C$.

F.2 DNA extraction and bacterial 16S rRNA amplicon sequencing

One aliquot each of 1,021 NPAs was used for 16S rRNA microbiome profiling. These include 487 out of 561 healthy NPAs collected from visits at 2 months, 6 months and 12 months of age; 380 out of 381 LRIs reported during the same period; and 154 out of 782 URIs (random selection of 0-2 URIs per infant). Overall, 397 healthy NPAs, 326 LRIs and 101 URIs were profiled for both virus and bacteria.

Total DNA was extracted and sequenced by MiSeq sequencing, where the amplicons were prepared using primers spanning the V4 region of the 16S rRNA gene and

containing barcoded reverse primers as published by Caporaso *et al.* [181].

Amplification of each sample was performed in quadruplicate to obtain enough amplicon for sequencing. Each plate also included a positive control (gDNA from *S. enterica* strain LT2, a bacterium not normally associated with the respiratory system (ATCC#700720D-5, USA)), and water as a negative control.

Due to the high throughput nature of this study, the sample DNA that was added into each PCR reaction was not quantified and normalized. Batch effects were instead mitigated by using a fixed volume (4 μL) of DNA template per well.

F.3 Quality control and taxonomic assignment

Quadruplicate sample amplicons were combined into a single well on a polymerase chain reaction (PCR) plate, then transferred to a fresh round-bottom polystyrene plate where they were purified. Amplicon quantitation was performed using the Quant-iT PicoGreen dsDNA quantitation kit (Life Technologies, Victoria, Australia) and fluorescence was determined on a Wallac Victor3 Multilabel counter (Perkin Elmer). PCR samples were equalized to 2 nM concentration (a neat aliquot was used where a sample fell below this concentration) and pools of 48, 60 or 96 barcoded samples were generated and sent for sequencing.

Paired end reads were merged and then quality filtered. A total of 33 million sequences were filtered out (13%), leaving 219 million for analysis. Sequences of sufficient quality were assigned to Operational Taxonomic Units (OTUs) using the closed reference method in QIIME v1.8 [182] with the Greengenes 99% OTU reference set, version 13_5 [](54). This reference set consists of more than 200,000 representative sequences obtained from clustering all sequences from the Greengenes reference database at 99% sequence similarity. Briefly, the closed reference OTU picking uses UCLUST [183] to search each sequence against the reference set, and assigns the sequence to an OTU based on the best hit at $\geq$ 99% sequence identity. Sequences that did not match the reference database (3%) were excluded from the analysis. This left an average of more than 200,000 (interquartile range: 108,000 - 255,000; 8 samples with less than 1000

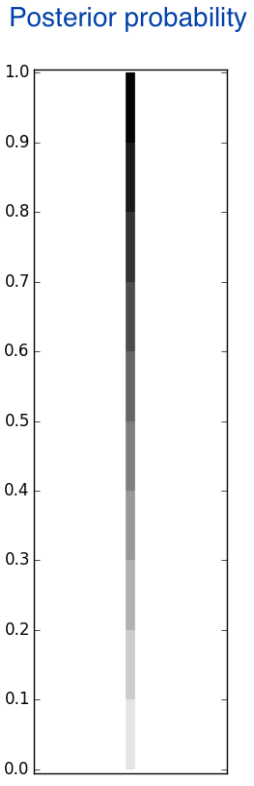
reads) taxonomy-assigned sequences for each NPA.

The relative abundance of each OTU was calculated for each NPA (i.e. reads matching the OTU, divided by total taxonomy-assigned reads for that sample). Most analyses were summarised at *genus* level, whereby all OTUs assigned to the same *genus* were collapsed into a single group for reporting the *genera* relative abundances.

Appendix G

Posterior legend

Legend showing correspondence between the shade with which an edge is rendered and its posterior probability in the given DBN. Black edges have near-unit probability, grey edges have intermediate probability, while the highly improbable have a "ghostly" appearance.



Appendix H

Mean-imputed DBNs with infection

Most of our DBNs were inferred after removing records with missing data in any of the variables. While this generally worked well, it was problematic for data in which some *years-of-life* had a lot more missing data than the other *years-of-life*. The most important example was infection data, for which the fourth and fifth *years-of-life* had a lot more missing data than the first three. Over the next few pages we present alternative DBNs for which missing data was mean-imputed instead of discarding incomplete records.

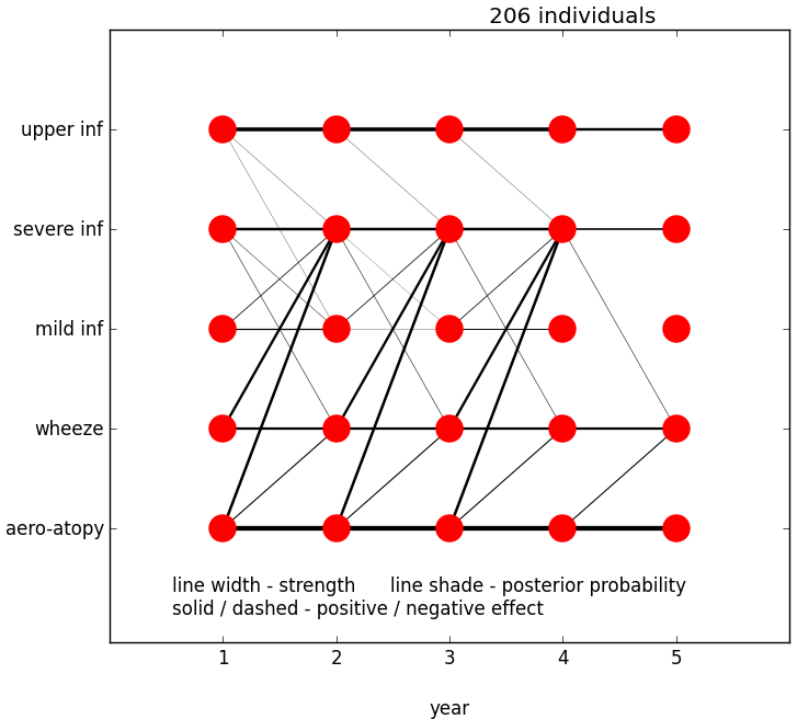


Figure H.1: ***Airborne-atopy*** **and** ***wheeze*** **each led to** ***severe-LRI*** **but not to** ***URI*** **or other** ***LRI*****. Neither** ***URI*** **nor** ***LRI*** **led to** ***airborne-atopy*** **nor** ***wheeze*****:**

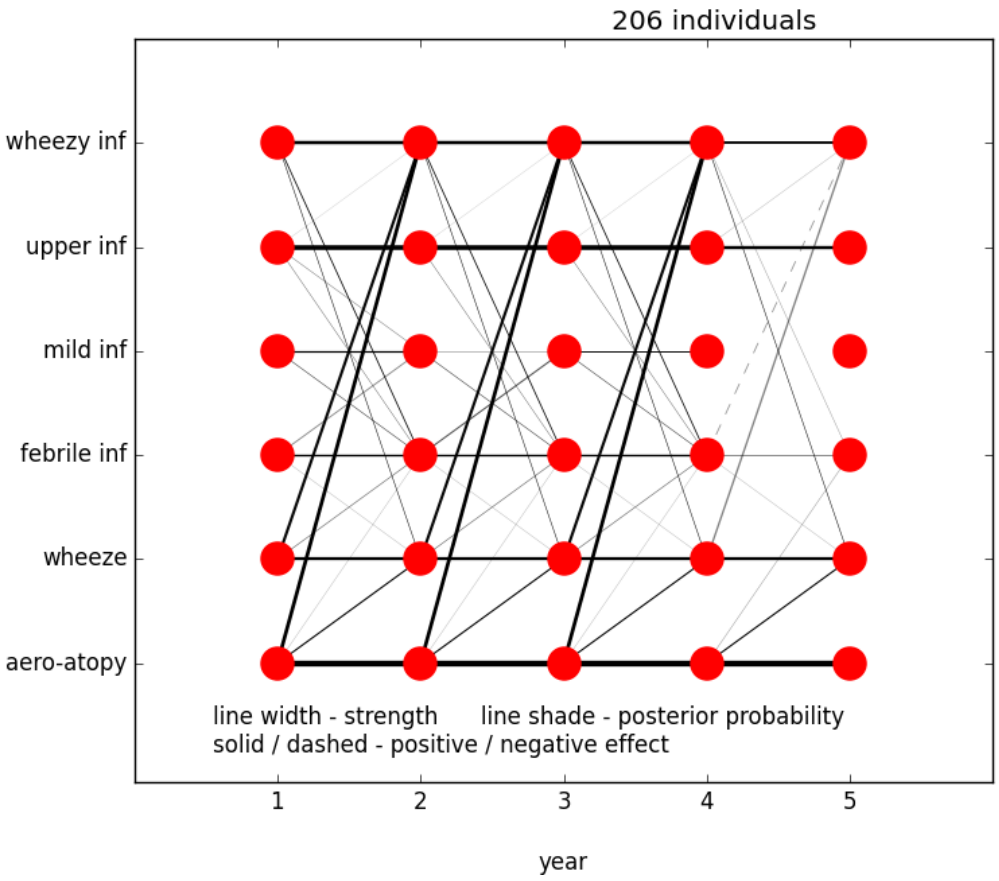


Figure H.2: ***Wheeze* and *airborne-atopy* led to *wheezy-LRI*. There was no negative link to *URI*:**

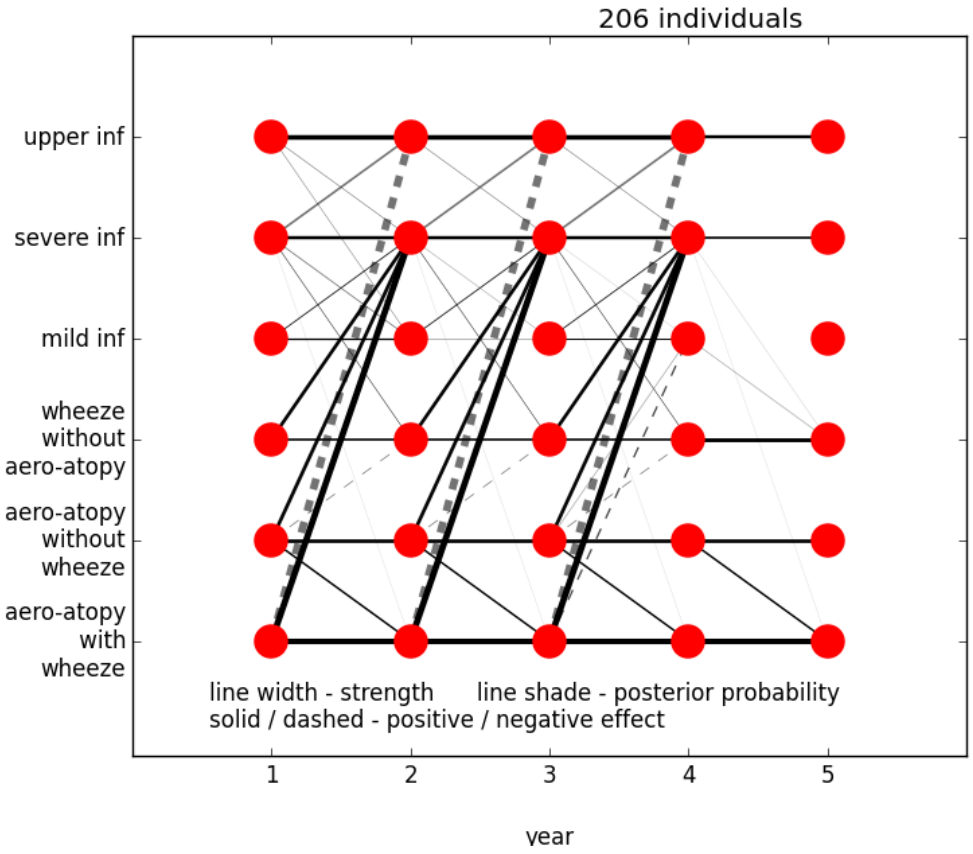


Figure H.3: **All *atopy* and *wheeze* led to *severe-LRI*, *aeroatopic-wheeze* most strongly. *Aeroatopic-wheeze* also led against *URI*, but with a lower posterior than in figure 5.9:**

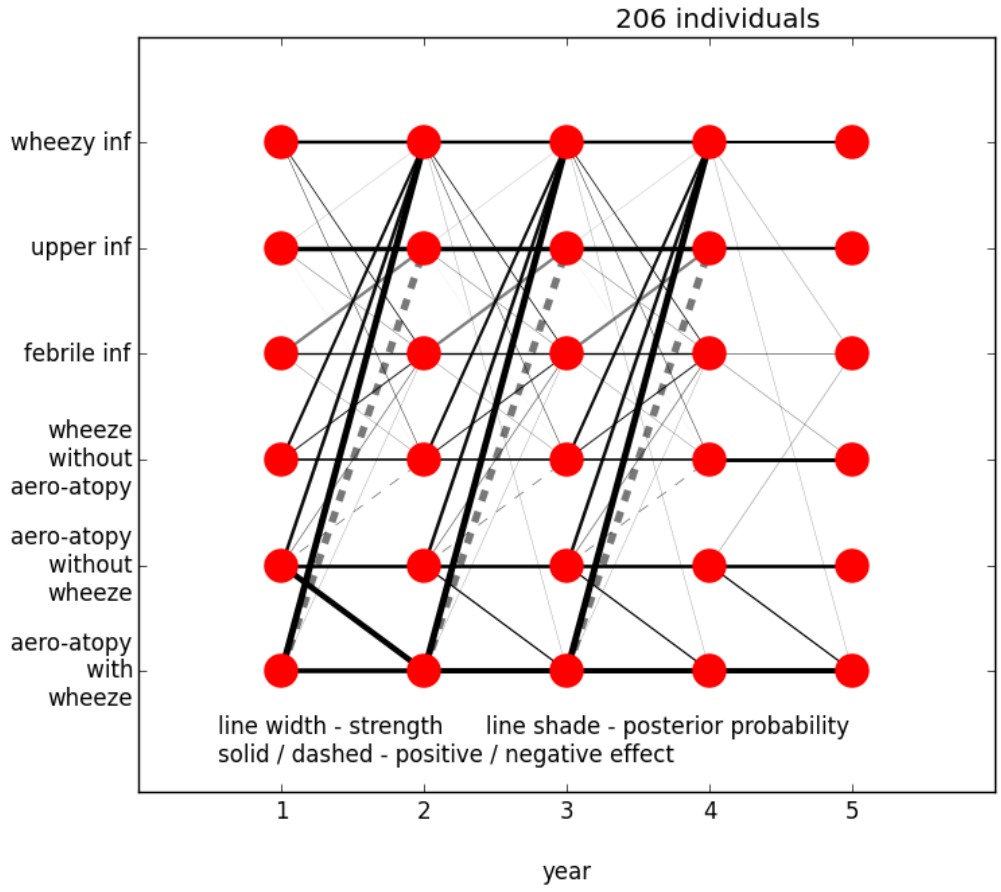


Figure H.4: ***Airborne-atopy*** **and** ***wheeze*** **lead to** ***wheezy-LRI*****. The two together have the strongest effect, and also lead away from** ***URI*****, indicating that they assist infectious agents in penetrating the lungs:**

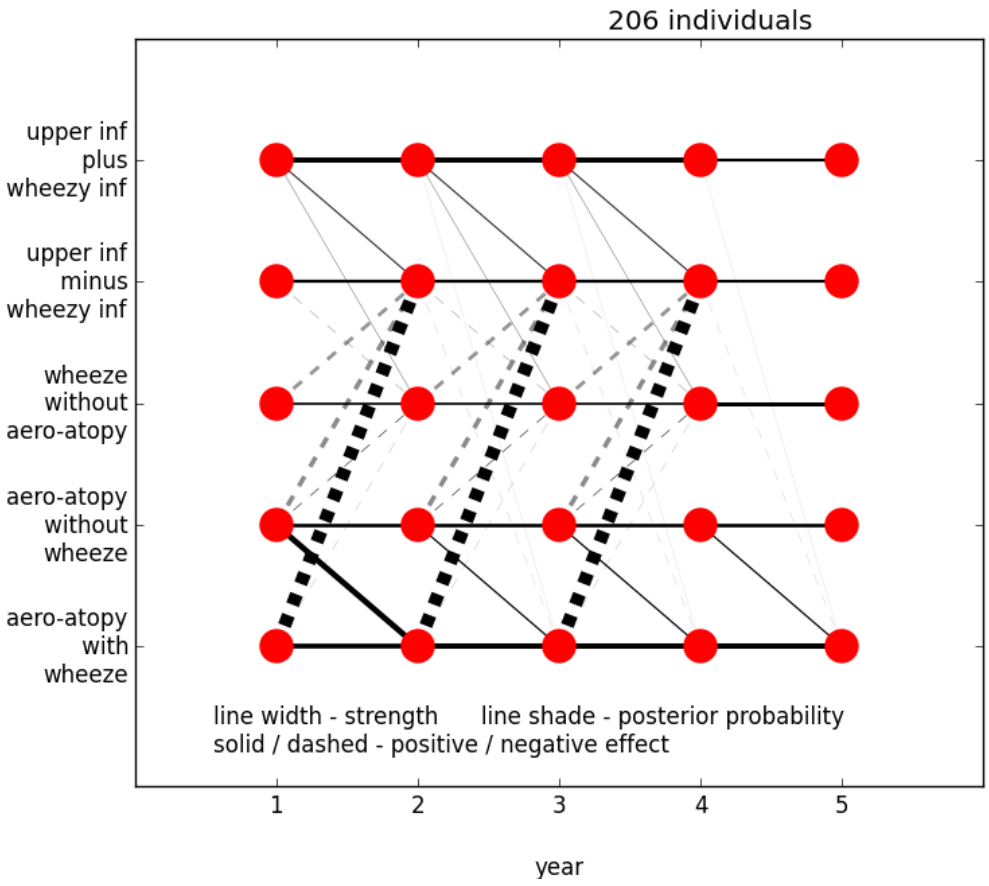


Figure H.5: ***Airborne-atopy* with *wheeze*, and to a lesser extent without *wheeze*, led to *URI* becoming *wheezy-LRI*:**

Appendix I

Other supplementary material

I.1 *Mild-viral-LRI* and the number of airborne *atopic* allergens

Figure 5.24 indicates a negative link from *mild-viral-LRI* in the first *year-of-life* to *aeroatopy-number*, but only when the *mild-viral-LRI* are accompanied by *airborne-atopy*, and this was supported by the χ-squared test, as shown in table 5.7. We also confirmed the lack of a link in the absence of *airborne-atopy* with the χ-squared test and the results are shown here in table I.1.

	mild-viral-LRI		
aeroatopy-number	previous year	same year	following year
First *year-of-life*	NA	.1117	.2429
Second *year-of-life*	.1773	.01672	.8141
Third *year-of-life*	.862	.652	NA

Table I.1: **χ-squared test p-values of the interaction between *aeroatopy-number* and the number of *mild-viral-LRI* in the previous, same and following year, where the *mild-viral-LRI* are *not* accompanied by *airborne-atopy*:** With the exception of the second *year-of-life*, there is no evidence of a relationship between these variables, and even that second-year p-value is large compared to those in other tables. This finding is consistent with the network in figure 5.24

I.2 Separating *severe-viral-LRI* into *wheezy- and febrile- viral-LRI*

In figures 5.7 and 5.8 we observed that splitting *severe-LRI* into it *febrile-* and *wheezy-LRI*, gave an improved DBN clearly signalling the importance of *wheezy-LRI*. The DBNs for *severe-viral-LRI* did not follow the same progression, probably due to *viral* data being limited to the first three *years-of-life*.

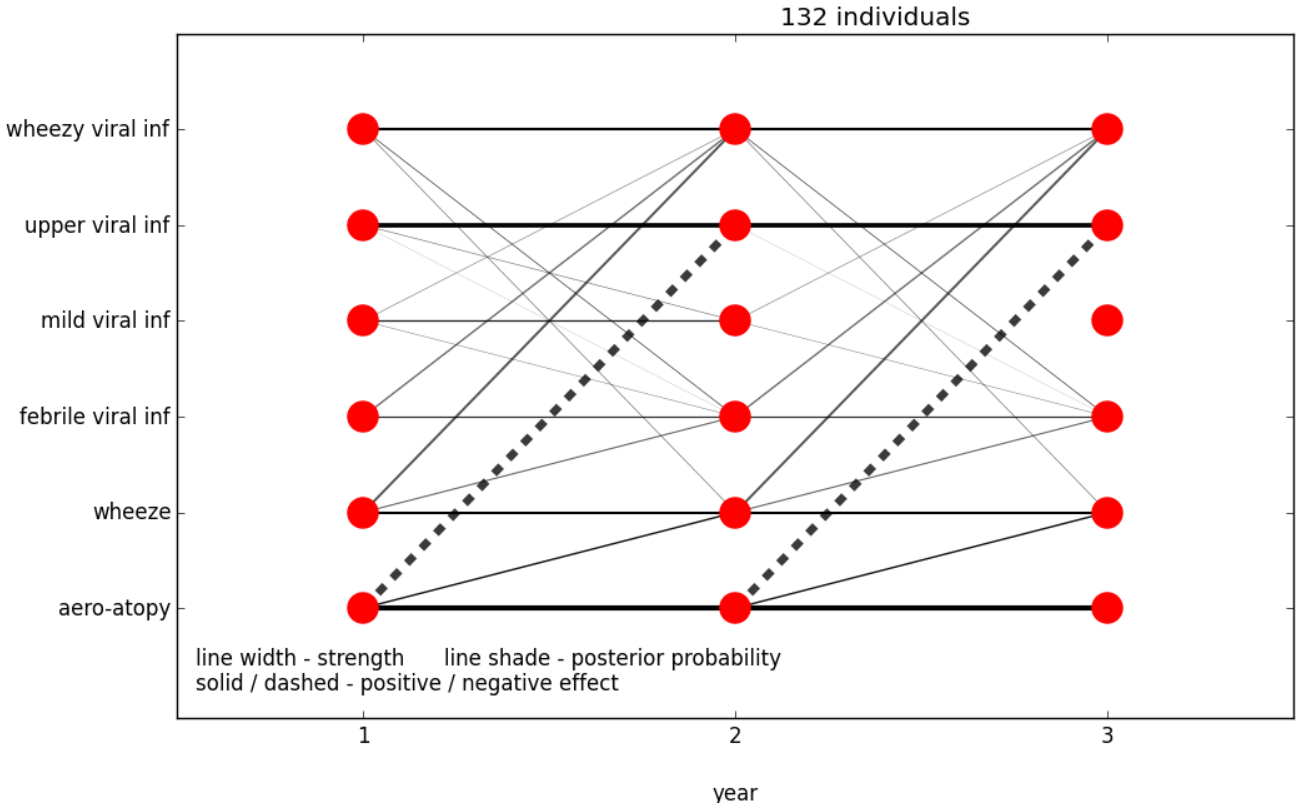


Figure I.1: ***Airborne-atopy* lead against *viral-URI*, with intermediate posterior:** Splitting *severe-viral-LRI* into its *febrile* and *wheezy* categories did not improve the network in the same way that splitting *severe-LRI* did (see figure 5.8).

I.3 *Aeroatopy-number* and *IgE* dynamics

Subsection 5.7.1 discussed interactions between *atopy-number* and (log of) *IgE* titres, based on the network in figure 5.29. The corresponding network (figure I.2), is very similar but of poorer quality.

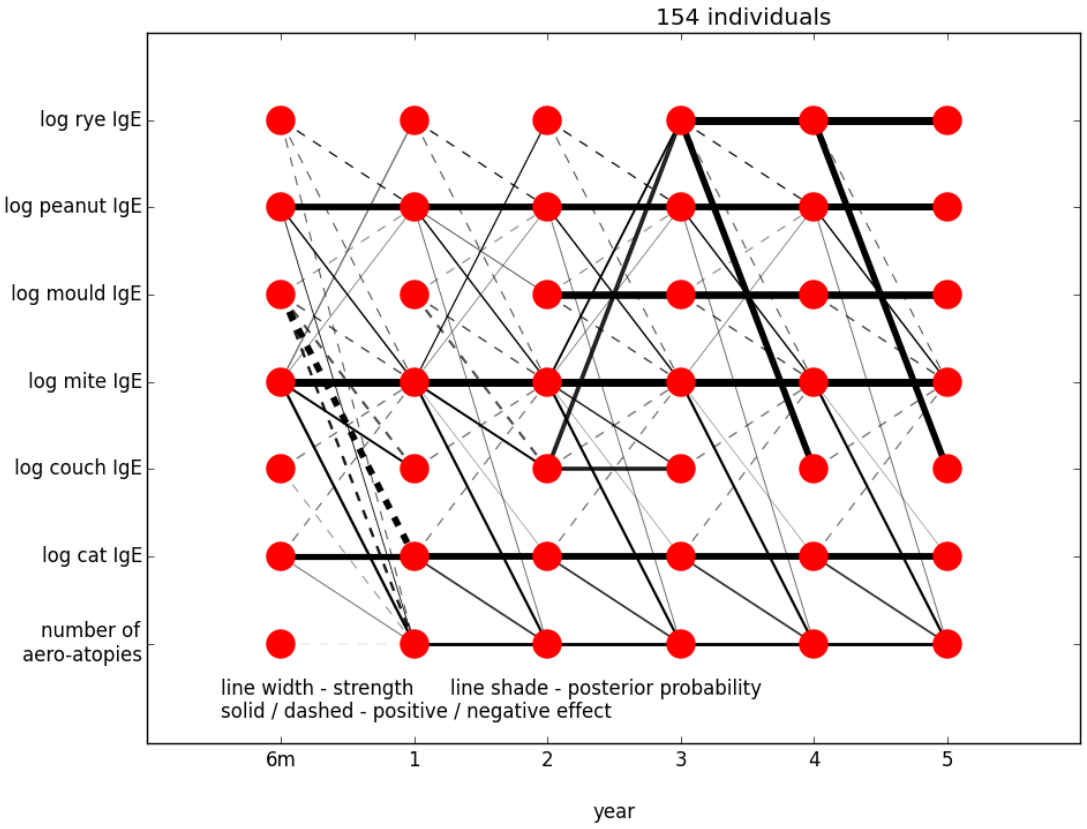


Figure I.2: **Interaction of *aeroatopy-number* (log of) *IgE*s:** Very similar to figure 5.29 but with a lot more faint cross-linking.

Appendix J

Commonly used terms

We present here some lists of the terms used commonly throughout this thesis.

J.1 Respiratory tract infections

Basic infections:

There is frequent reference to respiratory tract infections, which are variants on these two:

LRI lower respiratory tract infections, or the number thereof, during a given *year-of-life*.

URI upper respiratory tract infections, or the number thereof, during a given *year-of-life*. To be clear, this refers to repiratory tract infections which do not reach the lower respiratory tract.

Whether we are referring to the infections or to the number of them is clear from context.

Types of infection:

The *LRI*s, have many subtypes, described here:

febrile- the infections were accompanied by fever. CAS also applies this to *URI*, but we have not used *febrile-URI* in this thesis.

mild- the infections were accompanied by neither fever nor *wheeze*,

purely-febrile- the infections were accompanied by fever but not by *wheeze*,

purely-wheezy- the infections were accompanied by *wheeze* but not by fever,

severe- the infections were accompanied by fever, *wheeze*, or both,

wheezy- the infections were accompanied by *wheeze*.

Viral-status:

Swabs of the NP were taken during respiratory infections and tested for, among other things, the presence of *virii*. Which specific *virii* were present was determined by PCR analysis, but our work is only concerned with whether any virus was present. Both *URI*s and *LRI*s were sometimes accompanied by *virii*, denoted by:

viral- at least one virus was detected during the infections,

non-viral- no virus was detected during the infections.

Subtypes from these two lists can be taken in combination.

J.2 Immunity

Atopic-status

aeroallergen airborne allergen,

airborne-atopy *atopy* to any airborne allergen. The airborne allergens are *cat*, *mould*, *couch*, *rye* and *house-dust-mite* (binary), while *phadiatop* is a mixture of airborne allergens. The individual is then said to be *aeroatopic* to that aeroallergen (binary),

atopy having a blood concentration $\geq$ .35kU/L for any allergen-specific antibody. The individual is then said to be *atopic* to that allergen (binary),

nonairborne-atopy *atopy* where none of the allergens to which the individual is *atopic* are airborne (binary),

aeroatopy-number the number of airborne allergens to which the individual is *atopic*. This was not one of the CAS variables but derived from the various aeroallergen *IgE*s.

atopy-number the number of allergens to which the individual is *atopic*. This was not one of the CAS variables but derived from the various allergen *IgE*s.

IgE IgE antibodies to a given allergen, or the concentration thereof, according to context,

IL interleukin.

IFN-γ interferon-γ

Allergens

*IgE*s and interleukins are usually discussed in the context of these allergens:

cat cat,

couch couch grass,

house-dust-mite house-dust-mite,

mould mould,

peanut peanut,

phadiatop a broad mixture of airborne allergens,

rye rye grass,

ovalbumin ovalbumin,

tetanus tetanus.

J.3 Miscellaneous

***ARI* samples** Samples of the NP microbiome taken during the course of a respiratory infection,

***non-ARI* samples** Samples of the NP microbiome taken in the absence of a respiratory infection,

AUC Area Under the Curve, description in subsection 2.3.6,

BN Bayesian Network, described in subsection 2.2.1,

DBN Dynamic Bayesian Network, described in subsection 2.2.4,

wheeze whether a child had wheezed during a given *year-of-life*, as diagnosed by a doctor (binary),

transient wheeze whether *wheeze* was present during any of the first three *years-of-life* and absent in the fifth (binary),

year-of-life time periods were indicated according to the age of the individual and not from any given date. Similarly, first-year, second-year *etc.* are referring to the age of the individual and not to the age of the study. The corresponding understanding holds for *month-of-life* and *days-of-life*.

Bibliography

[1] M. Masoli, D. Fabian, S. Holt, R. Beasley, and G. I. for Asthma (GINA) Program, "The global burden of asthma: executive summary of the gina dissemination committee report," *Allergy*, vol. 59, no. 5, pp. 469–478, 2004.

[2] W. Eder, M. J. Ege, and E. von Mutius, "The asthma epidemic," *New England Journal of Medicine*, vol. 355, no. 21, pp. 2226–2235, 2006.

[3] W. C. Moore and S. P. Peters, "Severe asthma: an overview.," *J Allergy Clin Immunol*, vol. 117, no. 3, pp. 487–494, 2006.

[4] J. Bousquet, P. J. Bousquet, P. Godard, and J.-P. Daures, "The public health implications of asthma.," *Bull World Health Organ*, vol. 83, pp. 548–554, Jul 2005.

[5] M. M. H. Kusel, N. H. de Klerk, P. G. Holt, T. Kebadze, S. L. Johnston, and P. D. Sly, "Role of respiratory viruses in acute upper and lower respiratory tract illness in the first year of life: A birth cohort study," *The Pediatric Infectious Disease Journal*, vol. 25, no. 8, pp. 680–686, 2006.

[6] M. M. H. Kusel, N. H. de Klerk, T. Kebadze, V. Vohma, P. G. Holt, S. L. Johnston, and P. D. Sly, "Early-life respiratory viral infections, atopic sensitization, and risk of subsequent development of persistent asthma," *The Journal of allergy and clinical immunology*, vol. 119, no. 5, pp. 1105–1110, 2007.

[7] S. M. Teo, D. Mok, K. Pham, M. Kusel, M. Serralha, N. Troy, B. J. Holt, B. J. Hales, M. L. Walker, E. Hollams, Y. A. Bochkov, K. Grindle, S. L. Johnston, J. E. Gern, P. D. Sly, P. G. Holt, K. E. Holt, and M. Inouye, "The infant nasopharyngeal microbiome impacts severity of lower respiratory infection and risk of asthma development," *Cell Host & Microbe*, vol. 17, no. 5, pp. 704 – 715, 2015.

[8] S. Lebre, J. Becq, F. Devaux, M. Stumpf, and G. Lelandais, "Statistical inference of the time-varying structure of gene-regulation networks," *BMC Systems Biology*, vol. 4, no. 1, p. 130, 2010.

[9] J. H. Albert and S. Chib, "Bayesian analysis of binary and polychotomous response data," *Journal of the American statistical Association*, vol. 88, no. 422, pp. 669–679, 1993.

[10] R. T. Stein and F. D. Martinez, "Asthma phenotypes in childhood: lessons from an epidemiological approach," *Paediatric respiratory reviews*, vol. 5, pp. 155–161, 2004.

[11] S. Romanet-Manent, D. Charpin, A. Magnan, A. Lanteaume, and D. Vervloet, "Allergic vs nonallergic asthma: what makes the difference?," *Allergy*, vol. 57, no. 7, pp. 607–613, 2002.

[12] W. C. Moore, D. A. Meyers, S. E. Wenzel, W. G. Teague, H. Li, X. Li, R. D'Agostino, M. Castro, D. Curran-Everett, A. M. Fitzpatrick, B. Gaston, N. N. Jarjour, R. Sorkness, W. J. Calhoun, K. F. Chung, S. A. A. Comhair, R. A. Dweik, E. Israel, S. P. Peters, W. W. Busse, S. C. Erzurum, and E. R. Bleecker, "Identification of asthma phenotypes using cluster analysis in the severe asthma research program," *American Journal of Respiratory and Critical Care Medicine*, vol. 181, no. 4, pp. 315–323, 2010.

[13] E. M. Hollams, M. Deverell, M. Serralha, D. Suriyaarachchi, F. Parsons, G. Zhang, N. de Klerk, B. J. Holt, C. Ladyman, A. Sadowska, J. Rowe, R. Loh, P. D. Sly, and P. G. Holt, "Elucidation of asthma phenotypes in atopic teenagers through parallel immunophenotypic and clinical profiling," *Journal of Allergy and Clinical Immunology*, vol. 124, no. 3, pp. 463 – 470.e16, 2009.

[14] M. Pino-Yanes, A. Corrales, J. Cumplido, P. Poza, I. Sánchez-Machín, A. Sánchez-Palacios, J. Figueroa, O. Acosta-Fernández, N. Buset, J. C. García-Robaina, M. Hernández, J. Villar, T. Carrillo, and C. Flores, "Assessing the validity of asthma associations for eight candidate genes and age at diagnosis effects," *PLoS ONE*, vol. 8, no. 9, p. e73157, 2013.

[15] C. Miranda, A. Busacker, S. Balzar, J. Trudeau, and S. E. Wenzel, "Distinguishing severe asthma phenotypes: Role of age at onset and eosinophilic inflammation," *The Journal of allergy and clinical immunology*, vol. 113, pp. 101–108, 2004.

[16] P. G. Holt and P. D. Sly, "Viral infections and atopy in asthma pathogenesis: new rationales for asthma prevention and treatment," *Nat Med*, vol. 18, no. 5, pp. 726–735, 2012.

[17] W. W. Busse and R. F. Lemanske, "Asthma," *New England Journal of Medicine*, vol. 344, no. 5, pp. 350–362, 2001.

[18] O. V. Rossi, V. L. Kinnula, J. Tienari, and E. Huhti, "Association of severe asthma attacks with weather, pollen, and air pollutants.," *Thorax*, vol. 48, no. 3, pp. 244–248, 1993.

[19] P. G. Holt, J. Rowe, M. Kusel, F. Parsons, E. M. Hollams, A. Bosco, K. McKenna, L. Subrata, N. de Klerk, M. Serralha, B. J. Holt, G. Zhang, R. Loh, S. Ahlstedt, and P. D. Sly, "Toward improved prediction of risk for atopy and asthma among preschoolers: A prospective cohort study," *The Journal of allergy and clinical immunology*, vol. 125, no. 3, pp. 653–659.e7, 2010.

[20] P. D. Sly, A. L. Boner, B. BjÃűrksten, A. Bush, A. Custovic, P. A. Eigenmann, J. E. Gern, J. Gerritsen, E. Hamelmann, P. J. Helms, R. F. Lemanske, F. Martinez, S. Pedersen, H. Renz, H. Sampson, E. von Mutius, U. Wahn, and P. G. Holt, "Early identification of atopy in the prediction of persistent asthma in children," *The Lancet*, vol. 372, pp. 1100–1106, 2008.

[21] C. S. Ulrik, V. Backer, and A. Dirksen, "Mortality and decline in lung function in 213 adults with bronchial asthma: a ten-year follow up.," *J Asthma*, vol. 29, no. 1, pp. 29–38, 1992.

[22] C. Ulrik, V. Backer, A. Dirksen, M. Pedersen, and C. Koch, "Extrinsic and intrinsic asthma from childhood to adult age: a 10-yr follow-up," *Respiratory medicine*, vol. 89, no. 8, pp. 547–554, 1995.

[23] J.-C. Chang, H.-C. Kuo, T.-Y. Hsu, C.-Y. Ou, C.-A. Liu, H. Chuang, H.-M. Liang, H.-W. Huang, and K. D. Yang, "Different genetic associations of the ige production among fetus, infancy and childhood," *PLoS ONE*, vol. 8, no. 8, p. e70362, 2013.

[24] N. Pearce, N. AÃŕt-Khaled, R. Beasley, J. Mallol, U. Keil, E. Mitchell, and C. Robertson, "Worldwide trends in the prevalence of asthma symptoms: phase iii of the international study of asthma and allergies in childhood (isaac)," *Thorax*, vol. 62, no. 9, pp. 758–766, 2007.

[25] P.-O. Girodet, A. Ozier, I. Bara, J.-M. T. de Lara, R. Marthan, and P. Berger, "Airway remodeling in asthma: New mechanisms and potential for pharmacological intervention," *Pharmacology & Therapeutics*, vol. 130, no. 3, pp. 325 – 337, 2011.

[26] S. T. Holgate, "The airway epithelium is central to the pathogenesis of asthma.," *Allergol Int*, vol. 57, no. 1, pp. 1–10, 2008.

[27] C. M. Lloyd and S. Saglani, "Asthma and allergy: The emerging epithelium," *Nat Med*, vol. 16, no. 3, pp. 273–274, 2010.

[28] P. J. Barnes, "Intrinsic asthma: not so different from allergic asthma but driven by superantigens?," *Clinical & Experimental Allergy*, vol. 39, no. 8, pp. 1145–1151, 2009.

[29] P. D. Sly and F. Flack, "Susceptibility of children to environmental pollutants," *Annals of the New York Academy of Sciences*, vol. 1140, no. 1, pp. 163–183, 2008.

[30] C. R. Mackay, "Immunology: Dual personality of memory t cells," *Nature*, vol. 401, no. 6754, pp. 659–660, 1999.

[31] L. L. Cavanagh, R. Bonasio, I. B. Mazo, C. Halin, G. Cheng, A. W. M. van der Velden, A. Cariappa, C. Chase, P. Russell, M. N. Starnbach, P. A. Koni, S. Pillai, W. Weninger, and U. H. von Andrian, "Activation of bone marrow-resident memory t cells by circulating, antigen-bearing dendritic cells," *Nat Immunol*, vol. 6, no. 10, pp. 1029–1037, 2005.

[32] S. N. Abraham and A. L. St. John, "Mast cell-orchestrated immunity to pathogens," *Nat Rev Immunol*, vol. 10, no. 6, pp. 440–452, 2010.

[33] A. Berger, "Th1 and th2 responses: what are they?," *BMJ*, vol. 321, no. 7258, p. 424, 2000.

[34] W.-H. Boehncke, S. Boehncke, and M. P. Schön, "Managing comorbid disease in patients with psoriasis," *BMJ*, vol. 340, 2010.

[35] S. Illi, E. von Mutius, S. Lau, B. Niggemann, C. GrÃijber, and U. Wahn, "Perennial allergen sensitisation early in life and chronic asthma in children: a birth cohort study," *The Lancet*, vol. 368, pp. 763–770, 2006.

[36] H. L. Rhodes, P. Thomas, R. Sporik, S. T. Holgate, and J. J. Cogswell, "A birth cohort study of subjects at risk of atopy," *American Journal of Respiratory and Critical Care Medicine*, vol. 165, no. 2, pp. 176–180, 2002.

[37] D. Sherril, R. Stein, M. Kurzius-Spencer, and F. Martinez, "On early sensitization to allergens and development of respiratory symptoms," *Clinical & Experimental Allergy*, vol. 29, no. 7, pp. 905–911, 1999.

[38] V. O. Millien, W. Lu, J. Shaw, X. Yuan, G. Mak, L. Roberts, L. Z. Song, J. M. Knight, C. J. Creighton, A. Luong, F. Kheradmand, and D. B. Corry, "Cleavage of fibrinogen by proteinases elicits allergic responses through toll-like receptor 4," *Science*, vol. 341, no. 6147, pp. 792–796, 2013.

[39] S. Pérez-Rial, L. del Puerto-Nevado, R. Terrón-Expósito, A. Girón-Martínez, N. González-Mangado, and G. Peces-Barba, "Role of recently migrated monocytes in cigarette smoke-induced lung inflammation in different strain of mice," *PLoS ONE*, vol. 8, p. e72975, 09 2013.

[40] D. A. Hinds, G. McMahon, A. K. Kiefer, C. B. Do, N. Eriksson, D. M. Evans, B. St Pourcain, S. M. Ring, J. L. Mountain, U. Francke, G. Davey-Smith, N. J. Timpson, and J. Y. Tung, "A genome-wide association meta-analysis of self-reported allergy identifies shared and allergy-specific susceptibility loci," *Nat Genet*, vol. 45, pp. 907–911, 08 2013.

[41] C.-F. Liu, D. Drocourt, G. Puzo, J.-Y. Wang, and M. Riviere, "Innate immune response of alveolar macrophage to house dust mite allergen is mediated through tlr2/-4 co-activation," *PLoS ONE*, vol. 8, p. e75983, 10 2013.

[42] P. A. Frischmeyer-Guerrerio, A. L. Guerrerio, G. Oswald, K. Chichester, L. Myers, M. K. Halushka, M. Oliva-Hemker, R. A. Wood, and H. C. Dietz, "Tgfβ receptor mutations impose a strong predisposition for human allergic disease," *Science Translational Medicine*, vol. 5, no. 195, p. 195ra94, 2013.

[43] A. E. Semper, K. Heron, A. C. S. Woollard, J. P. Kochan, P. S. Friedmann, M. K. Church, and I. G. Reischl, "Surface expression of fcϵri on langerhans' cells of clinically uninvolved skin is associated with disease activity in atopic dermatitis, allergic asthma, and rhinitis," *Journal of Allergy and Clinical Immunology*, vol. 112, no. 2, pp. 411 – 419, 2003.

[44] J. Ker and T. V. Hartert, "The atopic march: what's the evidence?," *Annals of Allergy, Asthma & Immunology*, vol. 103, no. 4, pp. 282 – 289, 2009.

[45] S. Demehri, M. Morimoto, M. J. Holtzman, and R. Kopan, "Skin-derived tslp triggers progression from epidermal-barrier defects to asthma," *PLoS Biol*, vol. 7, p. e1000067, 05 2009.

[46] Z. Zhang, P. Hener, N. Frossard, S. Kato, D. Metzger, M. Li, and P. Chambon, "Thymic stromal lymphopoietin overproduced by keratinocytes in mouse skin aggravates experimental asthma," *Proceedings of the National Academy of Sciences*, vol. 106, no. 5, pp. 1536–1541, 2009.

[47] M. C. Siracusa, S. A. Saenz, D. A. Hill, B. S. Kim, M. B. Headley, T. A. Doering, E. J. Wherry, H. K. Jessup, L. A. Siegel, T. Kambayashi, E. C. Dudek, M. Kubo, A. Cianferoni, J. M. Spergel, S. F. Ziegler, M. R. Comeau, and D. Artis, "Tslp promotes interleukin-3-independent basophil haematopoiesis and type 2 inflammation," *Nature*, vol. 477, pp. 229–233, 09 2011.

[48] M.-P. F. Strippoli, B. D. Spycher, A. M. Pescatore, C. S. Beardsmore, M. Silverman, and C. E. Kuehni, "Viral wheezing is virus specific and not just host specific," *European Respiratory Journal*, vol. 39, no. 1, p. 229, 2012.

[49] W. W. Busse, "The relationship between viral infections and onset of allergic diseases and asthma," *Clinical & Experimental Allergy*, vol. 19, no. 1, pp. 1–9, 1989.

[50] W. H. Oddy, N. H. de Klerk, P. D. Sly, and P. G. Holt, "The effects of respiratory infections, atopy, and breastfeeding on childhood asthma," *European Respiratory Journal*, vol. 19, no. 5, pp. 899–905, 2002.

[51] S. Illi, E. von Mutius, S. Lau, R. Bergmann, B. Niggemann, C. Sommerfeld, and U. Wahn, "Early childhood infectious diseases and the development of asthma up to school age: a birth cohort study," *BMJ*, vol. 322, pp. 390–395, 2 2001.

[52] C. C. Copenhaver, J. E. Gern, Z. Li, P. A. Shult, L. A. Rosenthal, L. D. Mikus, C. J. Kirk, K. A. Roberg, E. L. Anderson, C. J. Tisler, D. F. DaSilva, H. J. Hiemke, K. Gentile, R. E. Gangnon, and R. F. Lemanske, "Cytokine response patterns, exposure to viruses, and respiratory infections in the first year of life," *American Journal of Respiratory and Critical Care Medicine*, vol. 170, no. 2, pp. 175–180, 2004.

[53] A. L. Wright, L. M. Taussig, C. G. Ray, H. R. Harrison, C. J. Holberg, and T. group health medical associates, "The tucson children's repiratory study: Ii. lower respiratory

tract illness in the first year of life," *American Journal of Epidemiology*, vol. 129, no. 6, pp. 1232–1246, 1989.

[54] M. K. Iwane, K. M. Edwards, P. G. Szilagyi, F. J. Walker, M. R. Griffin, G. A. Weinberg, C. Coulen, K. A. Poehling, L. P. Shone, S. Balter, C. B. Hall, D. D. Erdman, K. Wooten, B. Schwartz, and for the New Vaccine Surveillance Network, "Population-based surveillance for hospitalizations associated with respiratory syncytial virus, influenza virus, and parainfluenza viruses among young children," *Pediatrics*, vol. 113, no. 6, pp. 1758–1764, 2004.

[55] P. J. M. Openshaw and J. S. Tregoning, "Immune responses and disease enhancement during respiratory syncytial virus infection," *Clinical Microbiology Reviews*, vol. 18, no. 3, pp. 541–555, 2005.

[56] V. Noble, M. Murray, M. S. C. Webb, J. Alexander, A. S. Swarbrick, and A. D. Milner, "Respiratory status and allergy nine to 10 years after acute bronchiolitis," *Archives of Disease in Childhood*, vol. 76, no. 4, pp. 315–319, 1997.

[57] R. T. Stein, D. Sherrill, W. J. Morgan, C. J. Holberg, M. Halonen, L. M. Taussig, A. L. Wright, and F. D. Martinez, "Respiratory syncytial virus in early life and risk of wheeze and allergy by age 13 years," *The Lancet*, vol. 354, pp. 541–545, 1999.

[58] N. Sigurs, R. Bjarnason, F. Sigurbergsson, and B. Kjellman, "Respiratory syncytial virus bronchiolitis in infancy is an important risk factor for asthma and allergy at age 7," *American Journal of Respiratory and Critical Care Medicine*, vol. 161, pp. 1501–1507, 2013/12/02 2000.

[59] M. C. J. Kneyber, E. W. Steyerberg, R. de Groot, and H. A. Moll, "Long-term effects of respiratory syncytial virus (rsv) bronchiolitis in infants and young children: a quantitative review," *Acta PÃędiatrica*, vol. 89, no. 6, pp. 654–660, 2000.

[60] G. Wennergren and S. Kristjánsson, "Relationship between respiratory syncytial virus bronchiolitis and future obstructive airway diseases," *European Respiratory Journal*, vol. 18, no. 6, pp. 1044–1058, 2001.

[61] G. P. Rakes, E. Arruda, J. M. Ingram, G. E. Hoover, J. C. Zambrano, F. G. Hayden, T. A. E. Platts-Mills, and P. W. Heymann, "Rhinovirus and respiratory syncytial virus

in wheezing children requiring emergency care," *American Journal of Respiratory and Critical Care Medicine*, vol. 159, pp. 785–790, 2016/04/12 1999.

[62] S. L. Johnston, P. K. Pattemore, G. Sanderson, S. Smith, F. Lampe, L. Josephs, P. Symington, S. O. Toole, S. H. Myint, D. A. J. Tyrrell, and S. T. Holgate, "Community study of role of viral infections in exacerbations of asthma in 9-11 year old children," *BMJ*, vol. 310, pp. 1225–1229, 5 1995.

[63] J. E. G. W. W. Busse, "Viruses in asthma," *The Journal of allergy and clinical immunology*, vol. 100, pp. 147–150, 1997.

[64] J. A. McMillan, L. B. Weiner, A. M. Higgins, and K. MacKnight, "Rhinovirus infection associated with serious illness among pediatric patients," *The Pediatric Infectious Disease Journal*, vol. 12, no. 4, 1993.

[65] A. R. Smyth, R. L. Smyth, C. Y. Tong, C. A. Hart, and D. P. Heaf, "Effect of respiratory virus infections including rhinovirus on clinical status in cystic fibrosis.," *Archives of Disease in Childhood*, vol. 73, no. 2, pp. 117–120, 1995.

[66] D. J. Jackson, M. D. Evans, K. A. Roberg, E. L. Anderson, D. F. DaSilva, T. E. Pappas, R. E. G. C. J. Tisler, J. E. Gern, and R. F. Lemanske, "Allergic sensitization is a risk factor for rhinovirus wheezing illnesses during early childhood," *The Journal of allergy and clinical immunology*, vol. 125, 2010.

[67] J. P. Olenec, W. K. Kim, W.-M. Lee, F. Vang, T. E. Pappas, L. E. P. Salazar, M. D. Evans, J. Bork, K. Roberg, R. F. Lemanske, and J. E. Gern, "Weekly monitoring of children with asthma for infections and illness during common cold seasons," *The Journal of allergy and clinical immunology*, vol. 125, pp. 1001–1006.e1, 2010.

[68] F. D. Martinez, D. A. Stern, A. L. Wright, L. M. Taussig, and M. Halonen, "Differential immune responses to acute lower respiratory illness in early life and subsequent development of persistent wheezing and asthma," *The Journal of allergy and clinical immunology*, vol. 102, pp. 915–920, 1998.

[69] W. H. Oddy, P. G. Holt, P. D. Sly, A. W. Read, L. I. Landau, F. J. Stanley, G. E. Kendall, and P. R. Burton, "Association between breast feeding and asthma in 6 year old children: findings of a prospective birth cohort study," *BMJ*, vol. 319, pp. 815–819, 9 1999.

[70] P. G. Holt, J. W. Upham, and P. D. Sly, "Contemporaneous maturation of immunologic and respiratory functions during early childhood: Implications for development of asthma prevention strategies," *The Journal of allergy and clinical immunology*, vol. 116, pp. 16–24, 2005.

[71] F. J. Culley, A. M. J. Pennycook, J. S. Tregoning, T. Hussell, and P. J. M. Openshaw, "Differential chemokine expression following respiratory virus infection reflects th1- or th2-biased immunopathology," *Journal of Virology*, vol. 80, no. 9, pp. 4521–4527, 2006.

[72] D. J. Jackson, M. D. Evans, R. E. Gangnon, C. J. Tisler, T. E. Pappas, W.-M. Lee, J. E. Gern, and R. F. Lemanske, "Evidence for a causal relationship between allergic sensitization and rhinovirus wheezing in early life," *American Journal of Respiratory and Critical Care Medicine*, vol. 185, pp. 281–285, 2013/12/03 2012.

[73] M. A. Gill, A. K. Palucka, T. Barton, F. Ghaffar, H. Jafri, J. Banchereau, and O. Ramilo, "Mobilization of plasmacytoid and myeloid dendritic cells to mucosal sites in children with respiratory syncytial virus and other viral respiratory infections," *Journal of Infectious Diseases*, vol. 191, no. 7, pp. 1105–1115, 2005.

[74] A. S. McWilliam, A. M. Marsh, and P. G. Holt, "Inflammatory infiltration of the upper airway epithelium during sendai virus infection: involvement of epithelial dendritic cells.," *Journal of Virology*, vol. 71, no. 1, pp. 226–36, 1997.

[75] M. H. Grayson, D. Cheung, M. M. Rohlfing, R. Kitchens, D. E. Spiegel, J. Tucker, J. T. Battaile, Y. Alevy, L. Yan, E. Agapov, E. Y. Kim, and M. J. Holtzman, "Induction of high-affinity ige receptor on lung dendritic cells during viral infection leads to mucous cell metaplasia," *The Journal of Experimental Medicine*, vol. 204, no. 11, pp. 2759–2769, 2007.

[76] T. H. M. P. Consortium, "Structure, function and diversity of the healthy human microbiome," *Nature*, vol. 486, pp. 207–214, 06 2012.

[77] D. Bogaert, B. Keijser, S. Huse, J. Rossen, R. Veenhoven, E. van Gils, J. Bruin, R. Montijn, M. Bonten, and E. Sanders, "Variability and diversity of nasopharyngeal microbiota in children: A metagenomic analysis," *PLoS ONE*, vol. 6, p. e17035, 02 2011.

[78] J. J. Faith, J. L. Guruge, M. Charbonneau, S. Subramanian, H. Seedorf, A. L. Goodman, J. C. Clemente, R. Knight, A. C. Heath, R. L. Leibel, M. Rosenbaum, and J. I. Gordon, "The long-term stability of the human gut microbiota," *Science*, vol. 341, no. 6141, 2013.

[79] H.-J. Wu, I. I. Ivanov, J. Darce, K. Hattori, T. Shima, Y. Umesaki, D. R. Littman, C. Benoist, and D. Mathis, "Gut-residing segmented filamentous bacteria drive autoimmune arthritis via t helper 17 cells," *Immunity*, vol. 32, pp. 815–827, 06 2010.

[80] C. G. Buffie and E. G. Pamer, "Microbiota-mediated colonization resistance against intestinal pathogens," *Nat Rev Immunol*, vol. 13, pp. 790–801, 11 2013.

[81] V. K. Ridaura, J. J. Faith, F. E. Rey, J. Cheng, A. E. Duncan, A. L. Kau, N. W. Griffin, V. Lombard, B. Henrissat, J. R. Bain, M. J. Muehlbauer, O. Ilkayeva, C. F. Semenkovich, K. Funai, D. K. Hayashi, B. J. Lyle, M. C. Martini, L. K. Ursell, J. C. Clemente, W. van Treuren, W. A. Walters, R. Knight, C. B. Newgard, A. C. Heath, and J. I. Gordon, "Gut microbiota from twins discordant for obesity modulate metabolism in mice," *Science*, vol. 341, no. 6150, 2013.

[82] K. E. Fujimura, T. Demoor, M. Rauch, A. A. Faruqi, S. Jang, C. C. Johnson, H. A. Boushey, E. Zoratti, D. Ownby, N. W. Lukacs, and S. V. Lynch, "House dust exposure mediates gut microbiome *Lactobacillus* enrichment and airway immune defense against allergens and virus infection," *Proceedings of the National Academy of Sciences*, 2013.

[83] M. Kalliomäki and E. Isolauri, "Role of intestinal flora in the development of allergy," *Current Opinion in Allergy and Clinical Immunology*, vol. 3, no. 1, 2003.

[84] E. Y. Hsiao, S. W. McBride, S. Hsien, G. Sharon, E. R. Hyde, T. McCue, J. A. Codelli, J. Chow, S. E. Reisman, J. F. Petrosino, P. H. Patterson, and S. K. Mazmanian, "Microbiota modulate behavioral and physiological abnormalities associated with neurodevelopmental disorders," *Cell*, no. 0, pp. –, 2013.

[85] E. von Mutius, "Of attraction and rejection âĂŤ asthma and the microbial world," *New England Journal of Medicine*, vol. 357, no. 15, pp. 1545–1547, 2007. PMID: 17928604.

[86] H. Bisgaard, M. N. Hermansen, F. Buchvald, L. Loland, L. B. Halkjaer, K. BÃÿnnelykke, M. Brasholt, A. Heltberg, N. H. Vissing, S. V. Thorsen, M. Stage, and C. B. Pipper, "Childhood asthma after bacterial colonization of the airway in neonates," *New England Journal of Medicine*, vol. 357, no. 15, pp. 1487–1495, 2007. PMID: 17928596.

[87] B. J. Hales, L. J. Pearce, M. M. H. Kusel, P. G. Holt, P. D. Sly, and W. R. Thomas, "Differences in the antibody response to a mucosal bacterial antigen between allergic and non-allergic subjectssmoke-free legislation reduces exposure in children," *Thorax*, vol. 63, no. 3, pp. 221–227, 2008.

[88] B. J. Hales, L. Y. Chai, C. E. Elliot, L. J. Pearce, G. Zhang, T. K. Heinrich, W.-A. Smith, M. M. Kusel, P. G. Holt, P. D. Sly, and W. R. Thomas, "Antibacterial antibody responses associated with the development of asthma in house dust mite-sensitised and non-sensitised children," *Thorax*, vol. 67, no. 4, pp. 321–327, 2012.

[89] A. C. M. B. J. Hales, L. J. Pearce, I. A. Laing, C. M. Hayden, J. Goldblatt, P. N. L. SouÃńf, and W. R. Thomas, "Ige and igg antiâĂŞhouse dust mite specificities in allergic disease," *The Journal of allergy and clinical immunology*, vol. 118, pp. 361–367, 2006.

[90] E. M. Hollams, B. J. Hales, C. Bachert, W. Huvenne, F. Parsons, N. H. de Klerk, M. Serralha, B. J. Holt, S. Ahlstedt, W. R. Thomas, P. D. Sly, and P. G. Holt, "Th2-associated immunity to bacteria in teenagers and susceptibility to asthma," *European Respiratory Journal*, vol. 36, no. 3, pp. 509–516, 2010.

[91] K. K. Patel and W. C. Webley, "Evidence of infectious asthma phenotype: *Chlamydia*-induced allergy and pathogen-specific ige in a neonatal mouse model," *PLoS ONE*, vol. 8, p. e83453, 12 2013.

[92] H. Park, J. W. Shin, S.-G. Park, and W. Kim, "Microbial communities in the upper respiratory tract of patients with asthma and chronic obstructive pulmonary disease," *PLoS ONE*, vol. 9, p. e109710, 10 2014.

[93] C. Munck, J. Helby, C. G. Westergaard, C. Porsbjerg, V. Backer, and L. H. Hansen, "Smoking cessation and the microbiome in induced sputum samples from cigarette smoking asthma patients," *PLoS ONE*, vol. 11, pp. 1–11, 07 2016.

[94] Y. J. Huang, C. E. Nelson, E. L. Brodie, T. Z. DeSantis, M. S. Baek, J. Liu, T. Woyke, M. Allgaier, J. Bristow, J. P. Wiener-Kronish, E. R. Sutherland, T. S. King, N. Icitovic, R. J. Martin, W. J. Calhoun, M. Castro, L. C. Denlinger, E. DiMango, M. Kraft, S. P. Peters, S. I. Wasserman, M. E. Wechsler, H. A. Boushey, and S. V. Lynch, "Airway microbiota and bronchial hyperresponsiveness in patients with sub-optimally controlled

asthma," *The Journal of allergy and clinical immunology*, vol. 127, pp. 372–381.e3, 02 2011.

[95] R. P. Dickson and G. B. Huffnagle, "The lung microbiome: New principles for respiratory bacteriology in health and disease," *PLoS Pathog*, vol. 11, p. e1004923, 07 2015.

[96] T. T. Hansel, S. L. Johnston, and P. J. Openshaw, "Microbes and mucosal immune responses in asthma," *The Lancet*, vol. 381, no. 9869, pp. 861 – 873, 2013.

[97] Q. Zhang, M. Cox, Z. Liang, F. Brinkmann, P. A. Cardenas, R. Duff, P. Bhavsar, W. Cookson, M. Moffatt, and K. F. Chung, "Airway microbiota in severe asthma and relationship to asthma severity and phenotypes," *PLoS ONE*, vol. 11, pp. 1–16, 04 2016.

[98] J. E. Gern, C. M. Visness, P. J. Gergen, R. A. Wood, G. R. Bloomberg, G. T. O'Connor, M. Kattan, H. A. Sampson, F. R. Witter, M. T. Sandel, *et al.*, "The urban environment and childhood asthma (ureca) birth cohort study: design, methods, and study population," *BMC pulmonary medicine*, vol. 9, no. 1, p. 17, 2009.

[99] A. Malhotra, "Saturated fat is not the major issue," *BMJ*, vol. 347, 10 2013.

[100] J. Pearl and T. S. Verma, "A theory of inferred causation," in *Logic, Methodology and Philosophy of Science IX Proceedings of the Ninth International Congress of Logic, Methodology and Philosophy of Science* (B. S. Dag Prawitz and D. Westerståhl, eds.), vol. 134 of *Studies in Logic and the Foundations of Mathematics*, pp. 789 – 811, Amsterdam ; New York : Elsevier, 1995.

[101] J. Pearl, *Causality: models, reasoning and inference*, vol. 29. Cambridge Univ Press, 2000.

[102] N. Friedman, "Inferring cellular networks using probabilistic graphical models," *Science*, vol. 303, no. 5659, pp. 799–805, 2004.

[103] D. Madigan, J. York, and D. Allard, "Bayesian graphical models for discrete data," *International Statistical Review/Revue Internationale de Statistique*, vol. 63, pp. 215–232, 1995.

[104] M. B. Sesen, A. E. Nicholson, R. Banares-Alcantara, T. Kadir, and M. Brady, "Bayesian networks for clinical decision support in lung cancer care," *PLoS ONE*, vol. 8, p. e82349, 12 2013.

[105] C. S. Wallace and P. R. Freeman, "Estimation and inference by compact coding," *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 240–265, 1987.

[106] C. Wallace, K. B. Korb, and H. Dai, "Causal discovery via mml," in *ICML*, vol. 96, pp. 516–524, Citeseer, 1996.

[107] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the em algorithm," *Journal of the Royal Statistical Society*, vol. 39, no. 1, pp. 1–38, 1977.

[108] G. F. Cooper and E. Herskovits, "A bayesian method for the induction of probabilistic networks from data," *Machine Learning*, vol. 9, no. 4, pp. 309–347, 1992.

[109] B. A. Kidd, L. A. Peters, E. E. Schadt, and J. T. Dudley, "Unifying immunology with informatics and multiscale biology," *Nat Immunol*, vol. 15, pp. 118–127, 02 2014.

[110] T. Bayes, "An essay towards solving a problem in the doctrine of chances," *Phil. Trans. of the Royal Soc. of London*, vol. 53, pp. 370–418, 1763.

[111] A. Gelman and C. P. Robert, ""not only defended but also applied": The perceived absurdity of bayesian inference," *The American Statistician*, vol. 67, no. 1, pp. 1–5, 2013.

[112] D. Heckerman, "A tutorial on learning with bayesian networks," Technical Report MSR-TR-95-06, Microsoft Research, Advanced Technology Division, Microsoft Corporation, One Microsoft Way, Redmond, WA 98052, 1995.

[113] R. Kohavi, "Feature subset selection as search with probabilistic estimates," in *AAAI fall symposium on relevance*, vol. 224, 1994.

[114] R. Christensen, T. Hanson, and A. Jara, "Parametric nonparametric statistics," *The American Statistician*, vol. 62, no. 4, pp. 296–306, 2008.

[115] R. Kohavi and G. H. John, "Wrappers for feature subset selection," *Artificial Intelligence*, vol. 97, no. 1–2, pp. 273 – 324, 1997.

[116] G. H. John, R. Kohavi, K. Pfleger, *et al.*, "Irrelevant features and the subset selection problem.," in *ICML*, vol. 94, pp. 121–129, 1994.

[117] P. Langley and S. Sage, "Induction of selective bayesian classifiers," *CoRR*, vol. abs/1302.6828, 2013.

[118] P. R. Rosenbaum and D. B. Rubin, "The central role of the propensity score in observational studies for causal effects," *Biometrika*, vol. 70, no. 1, pp. 41–55, 1983.

[119] C. M. Zigler and F. Dominici, "Uncertainty in propensity score estimation: Bayesian methods for variable selection and model averaged causal effects," *Journal of the American Statistical Association*, vol. 0, no. ja, p. null, 0.

[120] T. Hoshino, "Semiparametric bayesian estimation for marginal parametric potential outcome modeling: Application to causal inference," *Journal of the American Statistical Association*, vol. 108, no. 504, pp. 1189–1204, 2013.

[121] E. Keogh and M. Pazzani, "Learning augmented bayesian classifiers: A comparison of distribution-based and classification-based approaches," in *Proceedings of the seventh international workshop on artificial intelligence and statistics*, pp. 225–230, Citeseer, 1999.

[122] N. Friedman and M. Goldszmidt, "Building classifiers using bayesian networks," in *Proceedings of the national conference on artificial intelligence*, pp. 1277–1284, 1996.

[123] N. Friedman, D. Geiger, and M. Goldszmidt, "Bayesian network classifiers," *Mach. Learn.*, vol. 29, pp. 131–163, Nov. 1997.

[124] J. Dougherty, R. Kohavi, and M. Sahami, "Supervised and unsupervised discretization of continuous features," in *ICML*, pp. 194–202, 1995.

[125] P. Domingos and M. Pazzani, "On the optimality of the simple bayesian classifier under zero-one loss," *Machine Learning*, vol. 29, no. 2-3, pp. 103–130, 1997.

[126] I. Rish, "An empirical study of the naive bayes classifier," in *IJCAI 2001 workshop on empirical methods in artificial intelligence*, vol. 3, pp. 41–46, 2001.

[127] P. Clark and T. Niblett, "The cn2 induction algorithm," *Machine learning*, vol. 3, no. 4, pp. 261–283, 1989.

[128] R. Hanson, J. Stutz, and P. Cheeseman, *Bayesian classification theory*. NASA Ames Research Center, Artificial Intelligence Research Branch, 1991.

[129] D. Draper and D. Fouskakis, "A case study of stochastic optimization in health policy: Problem formulation and preliminary results," *Journal of Global Optimization*, vol. 18,

pp. 399–416, 12 2000. Copyright - Kluwer Academic Publishers 2000; Last updated - 2010-06-05.

[130] D. D. Lewis, "Naive (bayes) at forty: The independence assumption in information retrieval," in *Machine Learning: ECML-98* (C. Nédellec and C. Rouveirol, eds.), vol. 1398 of *Lecture Notes in Computer Science*, pp. 4–15, Springer Berlin Heidelberg, 1998.

[131] P. Langley, W. Iba, and K. Thompson, "An analysis of bayesian classifiers," in *AAAI*, vol. 90, pp. 223–228, 1992.

[132] R. Blanco, I. Inza, M. Merino, J. Quiroga, and P. Larrañaga, "Feature selection in bayesian classifiers for the prognosis of survival of cirrhotic patients treated with {TIPS}," *Journal of Biomedical Informatics*, vol. 38, no. 5, pp. 376 – 388, 2005. Clinical Machine Learning.

[133] A. Yan, N. M. Laird, and C. Li, "Identifying rare variants using a bayesian regression approach," *BMC Proceedings*, vol. 5, no. 9, pp. 1–5, 2011.

[134] D. Heckerman and D. Geiger, "Likelihoods and parameter priors for bayesian networks," *Tech. MSRTR-95-54. Microsoft Research*, 1995.

[135] C. Rigaux, S. Ancelet, F. Carlin, C. Nguyen-thé, and I. Albert, "Inferring an augmented bayesian network to confront a complex quantitative microbial risk assessment model with durability studies: Application to bacillus cereus on a courgette purée production chain," *Risk Analysis*, vol. 33, no. 5, pp. 877–892, 2013.

[136] C. T. Perretti, S. B. Munch, and G. Sugihara, "Model-free forecasting outperforms the correct mechanistic model for simulated and experimental data," *Proceedings of the National Academy of Sciences*, vol. 110, no. 13, pp. 5253–5257, 2013.

[137] J. Smid, P. Volf, and G. Rao, "Monte carlo approach to bayesian regression modeling," in *Computer Intensive Methods in Control and Signal Processing* (M. Kárný and K. Warwick, eds.), pp. 169–180, Birkhäuser Boston, 1997.

[138] N. Metropolis and S. Ulam, "The monte carlo method," *Journal of the American Statistical Association*, vol. 44, no. 247, pp. 335–341, 1949.

[139] W. K. Hastings, "Monte carlo sampling methods using markov chains and their applications," *Biometrika*, vol. 57, no. 1, pp. 97–109, 1970.

[140] S. Geman and D. Geman, "Stochastic relaxation, gibbs distributions, and the bayesian restoration of images," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. PAMI-6, no. 6, pp. 721–741, 1984.

[141] C. Andrieu, N. de Freitas, A. Doucet, and M. I. Jordan, "An introduction to mcmc for machine learning," *Machine Learning*, vol. 50, no. 1-2, pp. 5–43, 2003.

[142] L. Tierney, "Markov chains for exploring posterior distributions," *the Annals of Statistics*, pp. 1701–1728, 1994.

[143] P. Dellaportas, J. J. Forster, and I. Ntzoufras, "On bayesian model and variable selection using mcmc," *Statistics and Computing*, vol. 12, no. 1, pp. 27–36, 2002.

[144] P. Carbonetto and M. Stephens, "Scalable variation inference for bayesian variable selection in regression, and its accuracy in genetic association studies," *Bayesian Analysis*, vol. 7, no. 1, pp. 73–108, 2012.

[145] M. I. Jordan, Z. Ghahramani, T. S. Jaakkola, and L. K. Saul, "An introduction to variational methods for graphical models," *Machine Learning*, vol. 37, no. 2, pp. 183–233, 1999.

[146] J. T. Ormerod and M. P. Wand, "Explaining variational approximations," *The American Statistician*, vol. 64, no. 2, pp. 140–153, 2010.

[147] M. J. Wainwright and M. I. Jordan, "Graphical models, exponential families, and variational inference," *Foundations and Trends® in Machine Learning*, vol. 1, no. 1-2, pp. 1–305, 2008.

[148] J. Yu, V. A. Smith, P. P. Wang, A. J. Hartemink, and E. D. Jarvis, "Advances to bayesian network inference for generating causal networks from observational biological data," *Bioinformatics*, vol. 20, no. 18, pp. 3594–3603, 2004.

[149] P. J. Green, "Reversible jump markov chain monte carlo computation and bayesian model determination," *Biometrika*, vol. 82, no. 4, pp. 711–732, 1995.

[150] M. Talih and N. Hengartner, "Structural learning with time-varying components: tracking the cross-section of financial time series," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 67, no. 3, pp. 321–341, 2005.

[151] M. Pagel and A. Meade, "Bayesian analysis of correlated evolution of discrete characters by reversible-jump markov chain monte carlo.," *The American Naturalist*, vol. 167, no. 6, pp. pp. 808–825, 2006.

[152] J. W. Robinson and A. J. Hartemink, "Non-stationary dynamic bayesian networks," in *Advances in Neural Information Processing Systems*, pp. 1369–1376, 2008.

[153] J. W. Robinson and A. J. Hartemink, "Learning non-stationary dynamic bayesian networks," *J. Mach. Learn. Res.*, vol. 11, pp. 3647–3680, 2010.

[154] U. Alon, "Network motifs: theory and experimental approaches," *Nat Rev Genet*, vol. 8, pp. 450–461, 06 2007.

[155] K. Sachs, O. Perez, D. Pe'er, D. A. Lauffenburger, and G. P. Nolan, "Causal protein-signaling networks derived from multiparameter single-cell data," *Science*, vol. 308, no. 5721, pp. 523–529, 2005.

[156] A. Fujita, J. R. Sato, H. M. Garay-Malpartida, P. A. Morettin, M. C. Sogayar, and C. E. Ferreira, "Time-varying modeling of gene expression regulatory networks using the wavelet dynamic vector autoregressive method," *Bioinformatics*, vol. 23, no. 13, pp. 1623–1630, 2007.

[157] A. Rao, A. O. Hero 3rd, D. J. States, and J. D. Engel, "Using directed information to build biologically relevant influence networks.," *Comput Syst Bioinformatics Conf*, vol. 6, pp. 145–156, 2007.

[158] C. Andrieu and A. Doucet, "Joint bayesian model selection and estimation of noisy sinusoids via reversible jump mcmc," *Signal Processing, IEEE Transactions on*, vol. 47, no. 10, pp. 2667–2676, 1999.

[159] A. Ng and A. Jordan, "On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes," *Advances in neural information processing systems*, vol. 14, p. 841, 2002.

[160] G. Bouchard, B. Triggs, *et al.*, "The tradeoff between generative and discriminative classifiers," in *16th IASC International Symposium on Computational Statistics (COMPSTAT'04)*, pp. 721–728, 2004.

[161] P. J. Bickel and E. Levina, "Some theory for fisher's linear discriminant function,'naive bayes', and some alternatives when there are many more variables than observations," *Bernoulli*, pp. 989–1010, 2004.

[162] J. Subramanian and R. Simon, "Overfitting in prediction models - is it a problem only in high dimensions?," *Contemporary Clinical Trials*, vol. 36, no. 2, pp. 636 – 641, 2013.

[163] J.-H. Xue and D. M. Titterington, "Comment on 'on discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes'," *Neural Processing Letters*, vol. 28, no. 3, pp. 169–187, 2008.

[164] J. A. Hanley and B. J. McNeil, "The meaning and use of the area under a receiver operating characteristic (roc) curve.," *Radiology*, vol. 143, no. 1, pp. 29–36, 1982.

[165] P. S. Carmack, J. S. Spence, and W. R. Schucany, "Generalised correlated cross-validation," *Journal of Nonparametric Statistics*, vol. 24, no. 2, pp. 269–282, 2012.

[166] R. Kohavi *et al.*, "A study of cross-validation and bootstrap for accuracy estimation and model selection," in *Ijcai*, vol. 14, pp. 1137–1145, 1995.

[167] J. Lötvall, C. A. Akdis, L. B. Bacharier, L. Bjermer, T. B. Casale, A. Custovic, R. F. Lemanske, A. J. Wardlaw, S. E. Wenzel, and P. A. Greenberger, "Asthma endotypes: a new approach to classification of disease entities within the asthma syndrome," *Journal of Allergy and Clinical Immunology*, vol. 127, no. 2, pp. 355–360, 2011.

[168] B. J. George, D. M. Reif, J. E. Gallagher, C. R. Williams-DeVane, B. L. Heidenfelder, E. E. Hudgens, W. Jones, L. Neas, E. A. C. Hubal, and S. W. Edwards, "Data-driven asthma endotypes defined from blood biomarker and gene expression data," *PLoS ONE*, vol. 10, p. e0117445, 02 2015.

[169] M. Raundhal, C. Morse, A. Khare, T. B. Oriss, J. Milosevic, J. Trudeau, R. Huff, J. Pilewski, F. Holguin, J. Kolls, S. Wenzel, P. Ray, and A. Ray, "High ifn-γ and low slpi mark severe asthma in mice and humans," *The Journal of Clinical Investigation*, vol. 125, pp. 0–0, 6 2015.

[170] B. D. Levy, P. J. Noel, M. M. Freemer, M. M. Cloutier, S. N. Georas, N. N. Jarjour, C. Ober, P. G. Woodruff, K. C. Barnes, B. G. Bender, *et al.*, "Future research directions

in asthma: An nhlbi working group report," *American Journal of Respiratory And Critical Care Medicine*, no. ja, 2015.

[171] K. M. Kloepfer, W. M. Lee, T. E. Pappas, T. J. Kang, R. F. Vrtis, M. D. Evans, R. E. Gangnon, Y. A. Bochkov, D. J. Jackson, R. F. Lemanske Jr, *et al.*, "Detection of pathogenic bacteria during rhinovirus infection is associated with increased respiratory symptoms and asthma exacerbations," *Journal of Allergy and Clinical Immunology*, vol. 133, no. 5, pp. 1301–1307, 2014.

[172] P. Burman, E. Chow, and D. Nolan, "A cross-validatory method for dependent data," *Biometrika*, vol. 81, no. 2, pp. 351–358, 1994.

[173] J. Bousquet, J. M. Anto, M. Wickman, T. Keil, R. Valenta, T. Haahtela, K. Lodrup Carlsen, M. van Hage, C. Akdis, C. Bachert, M. Akdis, C. Auffray, I. Annesi-Maesano, C. Bindslev-Jensen, A. Cambon-Thomsen, K. H. Carlsen, L. Chatzi, F. Forastiere, J. Garcia-Aymerich, U. Gehrig, S. Guerra, J. Heinrich, G. H. Koppelman, M. L. Kowalski, B. Lambrecht, C. Lupinek, D. Maier, E. MelÃľn, I. Momas, S. Palkonen, M. Pinart, D. Postma, V. Siroux, H. A. Smit, J. Sunyer, J. Wright, T. Zuberbier, S. H. Arshad, R. Nadif, C. Thijs, N. Andersson, A. Asarnoj, N. Ballardini, S. Ballereau, A. Bedbrook, M. Benet, A. Bergstrom, B. Brunekreef, E. Burte, M. Calderon, G. De Carlo, P. Demoly, E. Eller, M. P. Fantini, H. Hammad, C. Hohman, J. Just, M. Kerkhof, M. Kogevinas, I. Kull, S. Lau, N. Lemonnier, M. Mommers, M. Nawijn, A. Neubauer, S. Oddie, J. Pellet, I. Pin, D. Porta, Y. Saes, I. Skrindo, C. G. Tischer, M. Torrent, and L. von Hertzen, "Are allergic multimorbidities and ige polysensitization associated with the persistence or re-occurrence of foetal type 2 signalling? The MeDALL hypothesis," *Allergy*, vol. 70, no. 9, pp. 1062–1078, 2015.

[174] J. B. Kinney and G. S. Atwal, "Equitability, mutual information, and the maximal information coefficient," *Proceedings of the National Academy of Sciences*, 2014.

[175] R. Hanel, S. Thurner, and M. Gell-Mann, "How multiplicity determines entropy and the derivation of the maximum entropy principle for complex systems," *Proceedings of the National Academy of Sciences*, vol. 111, no. 19, pp. 6905–6910, 2014.

[176] W. Boomsma, J. Ferkinghoff-Borg, and K. Lindorff-Larsen, "Combining experiments and simulations using the maximum entropy principle," *PLoS Comput Biol*, vol. 10, p. e1003406, 02 2014.

[177] J. Zhao, Y. Zhou, X. Zhang, and L. Chen, "Part mutual information for quantifying direct associations in networks," *Proceedings of the National Academy of Sciences*, 2016.

[178] A. M. Singh, P. E. Moore, J. E. Gern, R. F. Lemanske Jr, and T. V. Hartert, "Bronchiolitis to asthma: a review and call for studies of gene–virus interactions in asthma causation," *American journal of respiratory and critical care medicine*, vol. 175, no. 2, pp. 108–119, 2007.

[179] P. Wu and T. V. Hartert, "Evidence for a causal relationship between respiratory syncytial virus infection and asthma," *Expert Review of Anti-infective Therapy*, vol. 9, no. 9, pp. 731–745, 2011.

[180] W. G. Waller and A. Jain, "On the monotonicity of the performance of bayesian classifiers (corresp.)," *Information Theory, IEEE Transactions on*, vol. 24, pp. 392–394, May 1978.

[181] J. G. Caporaso, C. L. Lauber, W. A. Walters, D. Berg-Lyons, J. Huntley, N. Fierer, S. M. Owens, J. Betley, L. Fraser, M. Bauer, *et al.*, "Ultra-high-throughput microbial community analysis on the illumina hiseq and miseq platforms," *The ISME journal*, vol. 6, no. 8, pp. 1621–1624, 2012.

[182] J. G. Caporaso, J. Kuczynski, J. Stombaugh, K. Bittinger, F. D. Bushman, E. K. Costello, N. Fierer, A. G. Peña, J. K. Goodrich, J. I. Gordon, *et al.*, "Qiime allows analysis of high-throughput community sequencing data," *Nature methods*, vol. 7, no. 5, pp. 335–336, 2010.

[183] R. C. Edgar, "Search and clustering orders of magnitude faster than blast," *Bioinformatics*, vol. 26, no. 19, pp. 2460–2461, 2010.